

# DECIDING THE DIMENSION OF EFFECTIVE DIMENSION REDUCTION SPACE FOR FUNCTIONAL AND HIGH-DIMENSIONAL DATA

BY YEHUA LI<sup>1</sup> AND TAIEN HSING<sup>2</sup>

*University of Georgia and University of Michigan*

In this paper, we consider regression models with a Hilbert-space-valued predictor and a scalar response, where the response depends on the predictor only through a finite number of projections. The linear subspace spanned by these projections is called the effective dimension reduction (EDR) space. To determine the dimensionality of the EDR space, we focus on the leading principal component scores of the predictor, and propose two sequential  $\chi^2$  testing procedures under the assumption that the predictor has an elliptically contoured distribution. We further extend these procedures and introduce a test that simultaneously takes into account a large number of principal component scores. The proposed procedures are supported by theory, validated by simulation studies, and illustrated by a real-data example. Our methods and theory are applicable to functional data and high-dimensional multivariate data.

**1. Introduction.** Li (1991) considered a regression model in which a scalar response depends on a multivariate predictor through an unknown number of linear projections, where the linear space spanned by the directions of the projections was named the effective dimension reduction (EDR) space of the model. Li (1991) introduced a  $\chi^2$  test to determine the dimension of the EDR space, and an estimation procedure, sliced inverse regression (SIR), to estimate the EDR space. Li's results focused on the case where  $p$ , the dimension of the predictor, is much smaller than  $n$ , the sample size. It is not obvious how to extend his results to high-dimensional multivariate data where  $p$  is comparable to or larger than  $n$ ; see Remark 5.4 in Li (1991).

---

Received December 2009.

<sup>1</sup>Supported by NSF Grant DMS-08-06131.

<sup>2</sup>Supported by NSF Grants DMS-08-08993 and DMS-08-06098.

*AMS 2000 subject classifications.* Primary 62J05; secondary 62G20, 62M20.

*Key words and phrases.* Adaptive Neyman test, dimension reduction, elliptically contoured distribution, functional data analysis, principal components.

This is an electronic reprint of the original article published by the Institute of Mathematical Statistics in *The Annals of Statistics*, 2010, Vol. 38, No. 5, 3028–3062. This reprint differs from the original in pagination and typographic detail.

Regression problems for functional data have drawn a lot of attention recently. In particular, regression models in which the predictor is functional while the response is scalar have been extensively investigated; for linear models, see Cardot, Ferraty and Sarda (2003), Ramsay and Silverman (2005), Cai and Hall (2006) and Crambes, Kneip and Sarda (2009); for nonlinear models, see Hastie and Tibshirani (1990), Cardot and Sarda (2005), James and Silverman (2005) and Müller and Stadtmüller (2005). Ferré and Yao (2003, 2005) extended SIR to a functional-data setting, and showed that the EDR space can be consistently estimated under regularity conditions provided that the true dimension of the space is known; see also Forzani and Cook (2007) and Ferré and Yao (2007). However, deciding the dimensionality of the EDR space is much more challenging in that case, and there has not been a formal procedure to date.

In this paper, we address the problem of deciding the dimensionality of the EDR space for both functional and high-dimensional multivariate data. As in Ferré and Yao (2003), we adopt the framework where the predictor takes value in an arbitrary Hilbert space. To better control the sample information in the (high-dimensional) predictor, we focus on the sample principal component scores rather than the raw data. Since the leading principal component scores optimally explain the variability in the predictor, it is natural to expect that the leading sample principal component scores also offer the most relevant information for the inference problem. Two statistical tests will be developed for testing whether the dimension of the EDR space is larger than a prescribed value; an estimator of the dimension of the EDR space will then be obtained by sequentially performing the tests developed. We will assume that the Hilbert-space-valued predictor has an elliptically contoured distribution, a common assumption for inverse regression problems [cf. Cook and Weisberg (1991), Schott (1994) and Ferré and Yao (2003)]. These tests will be first developed by focusing on a fixed number of principal component scores; it will be shown that the null distributions of the test statistics are asymptotically  $\chi^2$ . To address high and infinite-dimensional data, we propose an “adaptive Neyman” test, which combines the information in a sequence of  $\chi^2$  tests corresponding to an increasing number of principal component scores.

We introduce the background and notation in Section 2. The main theoretical results and test/estimation procedures are described in Section 3. Simulation studies are presented in Section 4, and a real application on near-infrared spectrum data is presented in Section 5. Finally, all of the proofs are collected in Appendix.

**2. Model assumptions and preliminaries.** Let  $\{X(t), t \in \mathcal{I}\}$  be a real-valued stochastic process with an index set  $\mathcal{I}$ . Assume that  $\mathbb{P}(X \in \mathcal{H}) = 1$ , where  $\mathcal{H}$  is some Hilbert space containing functions on  $\mathcal{I}$  and equipped

with inner product  $\langle \cdot, \cdot \rangle$ . We do not place any restriction on  $\mathcal{I}$  and so  $X$  can be extremely general. For instance, for multivariate data  $\mathcal{I}$  is a finite set, with  $p$  elements, say, and  $\mathcal{H}$  can then be taken as  $\mathbb{R}^p$  equipped with the usual dot product; in functional data analysis,  $\mathcal{H}$  is commonly assumed to be  $L^2(\mathcal{I})$  for some bounded interval  $\mathcal{I}$ , with inner product  $\langle g, h \rangle = \int_{\mathcal{I}} g(t)h(t) dt$ .

Consider the following multiple-index model:

$$(2.1) \quad Y = f(\langle \beta_1, X \rangle, \dots, \langle \beta_K, X \rangle, \varepsilon),$$

where  $Y$  is scalar,  $f$  is an arbitrary function,  $\beta_1, \dots, \beta_K$  are linearly independent elements in  $\mathcal{H}$ , and  $\varepsilon$  is a random error independent of  $X$ . Assume that  $f, K, \beta_1, \dots, \beta_K$  are all unknown, and we observe a random sample  $(X_i, Y_i), 1 \leq i \leq n$ , which are i.i.d. This is similar to the setting of Ferré and Yao (2003, 2005). Following Li (1991), we call  $\beta_1, \dots, \beta_K$  the EDR directions, and  $\text{span}(\beta_1, \dots, \beta_K)$  the EDR space. Without fixing  $f$ , the EDR directions are not identifiable; however, the EDR space is identifiable. The focus of this paper is the estimation of the dimension,  $K$ , of the EDR space.

We assume that the  $X_i$ 's are observed at each  $t \in \mathcal{I}$ . For functional data, this is an idealized assumption as no functions on a continuum can be fully observed. However, it is a reasonable approximation for densely observed smooth data, for which the Tecator data discussed in Section 5 is a good example. In that situation, for most theoretical and practical purposes, one can fit continuous curves to the discrete-time data and then treat the fitted curves as the true functional data; see, for example, Hall, Müller and Wang (2006), Cai and Hall (2006) and Zhang and Chen (2007). The case of sparsely observed functional data requires more attention and will not be studied in this paper. It may also be of interest to study the case where  $X$  contains measurement error; see (a) of Section 3.3.

*2.1. Principal components.* First, we focus on the generic process  $X$ . Denote the mean functions  $\mu$  of  $X$  by  $\mu(t) = \mathbb{E}\{X(t)\}, t \in \mathcal{I}$ . The covariance operator of  $X$  is the linear operator  $\Gamma_X := \mathbb{E}((X - \mu) \otimes (X - \mu))$ , where, for any  $h \in \mathcal{H}$ ,  $h \otimes h$  is the linear operator that maps any  $g \in \mathcal{H}$  to  $\langle h, g \rangle h$ . It can be seen that  $\Gamma_X$  is a well-defined compact operator so long as  $\mathbb{E}(\|X\|^4) < \infty$ , which we assume throughout the paper; see Eubank and Hsing (2010) for the mathematical details in constructing  $\mu$  and  $\Gamma_X$ . Then there exist nonnegative real numbers  $\omega_1 \geq \omega_2 \geq \dots$ , where  $\sum_j \omega_j < \infty$ , and orthonormal functions  $\psi_1, \psi_2, \dots$  in  $\mathcal{H}$  such that  $\Gamma_X \psi_j = \omega_j \psi_j$  for all  $j$ ; namely, the  $\omega_j$ 's are the eigenvalues and  $\psi_j$ 's the corresponding eigenfunctions of  $\Gamma_X$ . The  $\psi_j$ 's are commonly referred to as the principal components of  $X$ . It follows that

$$(2.2) \quad \Gamma_X = \sum_j \omega_j \psi_j \otimes \psi_j$$

and

$$(2.3) \quad X = \mu + \sum_j \xi_j \psi_j = \mu + \sum_j \sqrt{\omega_j} \eta_j \psi_j,$$

where the  $\xi_j$ 's are zero-mean, uncorrelated random variables with  $\text{Var}(\xi_j) = \omega_j$ , and the  $\eta_j$ 's are standardized  $\xi_j$ 's. Call  $\eta_j$  the standardized  $j$ th principal component score of  $X$ . The representations in (2.2) and (2.3) are commonly referred to as the principal component decomposition and the Karhunen–Loève expansion, respectively; see Ash and Gardner (1975) and Eubank and Hsing (2010) for details.

In view of (2.1) and (2.3), any component of  $\beta_k$  that is in the orthogonal complement of the span of the  $\psi_j$  is not estimable. As explained above, this paper does not address the estimation of the  $\beta_k$ . Thus, assume without generality that the  $\beta_k$ 's are spanned by the  $\psi_j$ 's and write

$$(2.4) \quad \beta_k = \sum_j \frac{b_{kj}}{\sqrt{\omega_j}} \psi_j.$$

By (2.3) and (2.4),  $\langle \beta_k, X \rangle = \langle \beta_k, \mu \rangle + \sum_j b_{kj} \eta_j$ , and (2.1) can be re-expressed as

$$Y = f\left(\sum_j b_{1j} \eta_j, \dots, \sum_j b_{Kj} \eta_j, \varepsilon\right),$$

where, for simplicity, the constants  $\langle \beta_1, \mu \rangle, \dots, \langle \beta_K, \mu \rangle$  are absorbed by  $f$ . For the i.i.d. sample  $(X_1, Y_1), \dots, (X_n, Y_n)$ , let  $\eta_{ij}$  be the standardized  $j$ th principal component score of  $X_i$ , and write

$$(2.5) \quad Y_i = f\left(\sum_j b_{1j} \eta_{ij}, \dots, \sum_j b_{Kj} \eta_{ij}, \varepsilon_i\right).$$

**2.2. Elliptically contoured distributions.** As mentioned in Section 1, the relevance of elliptical symmetry is evident in the inference of (2.1). We devote this subsection to a brief introduction of the notion of elliptically contoured distribution for Hilbert-space-valued variables.

Let  $X$  be as defined in Section 2.1. By the assumption  $\mathbb{E}(\|X\|^4) < \infty$ , the distribution of  $X$  is determined by the (marginal) distributions of the random variables  $\langle h, X \rangle, h \in \mathcal{H}$ . Say that  $X$  has an elliptically contoured distribution if

$$(2.6) \quad \mathbb{E}(e^{i\langle h, X - \mu \rangle}) = \phi(\langle h, \Sigma h \rangle), \quad h \in \mathcal{H},$$

for some function  $\phi$  on  $\mathbb{R}$  and self-adjoint, nonnegative operator  $\Sigma$ . Recall that  $X$  is said to be a Gaussian process if  $\langle h, X \rangle$  is normally distributed

for any  $h \in \mathcal{H}$ , and so (2.6) holds with  $\phi(t) = \exp(-t/2)$  and  $\Sigma = \Gamma_X$ . However, (2.6) in general describes a much larger class of distributions.

The mathematics necessary to characterize elliptically contoured distributions was worked out in Schoenberg (1938); see Cambanis, Huang and Simons (1981) and Li (2007). It follows that definition (2.6) implies that  $\Sigma$  is a constant multiple of  $\Gamma_X$  and  $\phi(t^2)$  is a characteristic function. More explicitly, (2.6) leads to the characterization

$$(2.7) \quad X - \mu \stackrel{d}{=} \Theta \check{X},$$

where  $\Theta$  and  $\check{X}$  are independent,  $\Theta$  is a nonnegative random variable with  $\mathbb{E}(\Theta^2) = 1$  and  $\check{X}$  has the same covariance operator as  $X$ ; if  $X \in \mathbb{R}^p$  and  $\text{rank}(\Gamma_X) = k \geq 1$  then  $\check{X} \stackrel{d}{=} \Theta A_{p \times k} U_{k \times 1}$  where  $AA^T = \Gamma_X$  and  $U$  is uniformly distributed on the  $k$ -dimensional sphere with radius  $\sqrt{k}$ ; if  $\text{rank}(\Gamma_X) = \infty$  then  $\check{X}$  is necessarily a zero-mean Gaussian process. Recall that  $U_{k \times 1}$  is asymptotically Gaussian [cf. Spruill (2007)] and so the infinite-dimensional representation can be viewed as the limit of the finite-dimensional one.

**2.3. Functional inverse regression.** To introduce functional inverse regression, we first state some conditions:

- (C1)  $\mathbb{E}(\|X\|^4) < \infty$ .
- (C2) For any function  $b \in \mathcal{H}$ , there exist some constants  $c_0, \dots, c_K$  such that

$$\mathbb{E}(\langle b, X \rangle | \langle \beta_1, X \rangle, \dots, \langle \beta_K, X \rangle) = c_0 + c_1 \langle \beta_1, X \rangle + \dots + c_K \langle \beta_K, X \rangle.$$

- (C3)  $X$  has an elliptically contoured distribution; namely, (2.7) holds.

Conditions (C1)–(C3) are standard conditions in the inverse regression literature; see, for instance, Ferré and Yao (2003, 2005). As mentioned earlier, condition (C1) guarantees the principal decomposition; moreover, it also ensures the convergence rate of  $n^{-1/2}$  in the estimation of the eigenvalues and eigenspaces of  $\Gamma_X$  based on an i.i.d. sample  $X_1, \dots, X_n$ ; see Dauxois, Pousse and Romain (1982). Condition (C2) is a direct extension of (3.1) in Li (1991) which addresses multivariate data. If  $X$  is a Gaussian process, then projections of  $X$  are jointly normal, from which (C2) follows easily. Condition (C3) describes a broader class of processes satisfying (C2) than the Gaussian process; for convenience (C3) is often assumed in lieu of (C2).

Call the collection  $\{\mathbb{E}(X(t)|Y), t \in \mathcal{J}\}$  of random variables the inverse regression process and denote its covariance operator by  $\Gamma_{X|Y}$ . We will use the notation  $\text{Im}(T)$ , for any operator  $T$ , to denote the range of  $T$ . The following result, first appeared in Ferré and Yao (2003), is a straightforward extension of Theorem 3.1 of Li (1991).

THEOREM 2.1. Under (C1) and (C2),  $\text{Im}(\Gamma_{X|Y}) \subset \text{span}(\Gamma_X \beta_1, \dots, \Gamma_X \beta_K)$ .

Theorem 2.1 implies that  $\text{span}(\Gamma_X \beta_1, \dots, \Gamma_X \beta_K)$  contains all of the eigenfunctions that correspond to the nonzero eigenvalues of  $\Gamma_{X|Y}$ . Consequently, if  $\Gamma_{X|Y}$  has  $K$  nonzero eigenvalues, then the space spanned by the eigenfunctions is precisely  $\text{span}(\Gamma_X \beta_1, \dots, \Gamma_X \beta_K)$ . In that case, one can in principle estimate  $\text{span}(\beta_1, \dots, \beta_K)$  through estimating both  $\Gamma_X$  and  $\Gamma_{X|Y}$ . This forms the basis for the estimation of the EDR space [cf. Li (1991) and Ferré and Yao (2003, 2005)].

While  $\Gamma_{X|Y}$  is finite-dimensional under (C1) and (C2), if  $\mathcal{H}$  is infinite-dimensional then its definition still involves infinite-dimensional random functions. In order to implement any inference procedure, we consider a finite-dimensional adaptation using principal components.

Let  $m$  be any positive integer, where  $m \leq n-1$  and, if  $X$  is  $p$ -dimensional,  $m \leq p$ . Define  $\mathbf{b}_{k,(m)} = (b_{k1}, \dots, b_{km})^T$ ,  $\boldsymbol{\eta}_{i,(m)} = (\eta_{i1}, \dots, \eta_{im})^T$  and  $\varsigma_{ik} = \sum_{j>m} \eta_{ij} b_{kj}$ . Then (2.5) can be expressed as

$$(2.8) \quad Y_i = f(\mathbf{b}_{1,(m)}^T \boldsymbol{\eta}_{i,(m)} + \varsigma_{i1}, \dots, \mathbf{b}_{K,(m)}^T \boldsymbol{\eta}_{i,(m)} + \varsigma_{iK}, \varepsilon_i).$$

If one regards the  $\boldsymbol{\eta}_{i,(m)}$  as predictors and combine the  $\varsigma_{ik}$  with  $\varepsilon_i$  to form the error, then (2.8) bears considerable similarity with the multivariate model of Li (1991). One fundamental difference is that although the  $\varsigma_{ik}$  are uncorrelated with  $\boldsymbol{\eta}_{i,(m)}$ , they might not be independent of  $\boldsymbol{\eta}_{i,(m)}$ , unless  $X$  is Gaussian. Another major difference is that we do not directly observe  $\boldsymbol{\eta}_{i,(m)}$  so that this model might be viewed as a variation of the errors-in-variables model in Carroll and Li (1992). Our estimator for  $K$  will be motivated by the finite-dimensional model (2.8). The details of the procedure, including the role of  $m$ , will be explained in Section 3. To pave the way for that, we briefly discuss the inference of the  $\mathbf{b}_{k,(m)}$  below.

We first need to estimate  $\boldsymbol{\eta}_{i,(m)}$ . Let

$$\bar{X} = n^{-1} \sum_{i=1}^n X_i \quad \text{and} \quad \hat{\Gamma}_X = n^{-1} \sum_{i=1}^n (X_i - \bar{X}) \otimes (X_i - \bar{X})$$

be the sample mean function and the sample covariance operator, respectively. Let  $\hat{\omega}_j$  and  $\hat{\psi}_j$  be the  $j$ th sample eigenvalue and eigenfunction of  $\hat{\Gamma}_X$ . By Dauxois, Pousse and Romain (1982),  $\hat{\omega}_j$  and  $\hat{\psi}_j$  are root- $n$  consistent under (C1). The standardized  $j$ th principal component scores of  $X_i$  are then estimated by  $\hat{\eta}_{ij} = \hat{\omega}_j^{-1/2} \langle \hat{\psi}_j, X_i - \bar{X} \rangle$ ; let  $\hat{\boldsymbol{\eta}}_{i,(m)} = (\hat{\eta}_{i1}, \dots, \hat{\eta}_{im})^T$ .

Based on the “data”  $(\hat{\boldsymbol{\eta}}_{i,(m)}, Y_i), 1 \leq i \leq n$ , the usual sliced inverse regression (SIR) algorithm can be carried out as follows. Partition the range of  $Y$  into disjoint intervals,  $S_h, h = 1, \dots, H$ , where  $p_h := \mathbb{P}(Y \in S_h) > 0$  for all  $h$ . Define

$$(2.9) \quad \vartheta_{j,h} = \mathbb{E}(\eta_j | Y \in S_h), \quad \boldsymbol{\vartheta}_{h,(m)} = \mathbb{E}(\boldsymbol{\eta}_{(m)} | Y \in S_h) = (\vartheta_{1,h}, \dots, \vartheta_{m,h})^T$$

and

$$(2.10) \quad \mu_h = \mathbb{E}(X|Y \in S_h) = \mu + \sum_j \omega_j^{1/2} \vartheta_{j,h} \psi_j.$$

Let  $V_{(m)} = \sum_h p_h \boldsymbol{\vartheta}_{h,(m)} \boldsymbol{\vartheta}_{h,(m)}^T$  be the between-slice covariance matrix. In the finite-dimensional model (2.8) with  $(\varsigma_{i1}, \dots, \varsigma_{iK}, \varepsilon_i)$  playing the role of error,  $V_{(m)}$  is the sliced-inverse-regression covariance matrix. The eigenvectors of  $V_{(m)}$  corresponding to the nonzero eigenvalues are contained in  $\text{span}(\mathbf{b}_{1,(m)}, \dots, \mathbf{b}_{K,(m)})$ . The matrix  $V_{(m)}$  is estimated by the corresponding sample version  $\hat{V}_{(m)} = \sum_{h=1}^H \hat{p}_h \hat{\boldsymbol{\vartheta}}_{h,(m)} (\hat{\boldsymbol{\vartheta}}_{h,(m)})^T$ , where

$$(2.11) \quad \begin{aligned} \hat{p}_h &= n_h/n, \quad n_h = \sum_i I(Y_i \in S_h) \quad \text{and} \\ \hat{\boldsymbol{\vartheta}}_{h,(m)} &= \frac{1}{n_h} \sum_{i=1}^n \hat{\boldsymbol{\eta}}_{i,(m)} I(Y_i \in S_h). \end{aligned}$$

Letting  $\hat{\mathbf{b}}_{1,(m)}, \dots, \hat{\mathbf{b}}_{K,(m)}$  be the first  $K$  eigenvectors of  $\hat{V}_{(m)}$ , the estimators of  $\beta_k$ 's are given by

$$\hat{\beta}_k(t) = \sum_{j=1}^m \hat{\omega}_j^{-1/2} \hat{b}_{kj} \hat{\psi}_j(t), \quad k = 1, \dots, K.$$

In order for  $\text{span}(\hat{\beta}_1, \dots, \hat{\beta}_K)$  to consistently estimate the EDR space, it is necessary that  $\text{span}(\mu_1 - \mu, \dots, \mu_h - \mu)$  have the same dimension as the EDR space, and that  $m$  tends to  $\infty$  with  $n$  in some manner. However, first and foremost, we must know  $K$  beforehand, which makes the determination of  $K$  a fundamental issue.

The matrix  $\hat{V}_{(m)}$  will be our basis for deciding  $K$ . Here, we define some notation related to  $V_{(m)}$  and  $\hat{V}_{(m)}$  for future use. For any  $m \times 1$  vector  $\mathbf{u}$ , let  $\mathcal{J}_{\mathbf{u}} = I - \mathbf{u}\mathbf{u}^T$ ; let  $\mathbf{g} = (g_1, \dots, g_H) = (p_1^{1/2}, \dots, p_H^{1/2})$ , and  $\hat{\mathbf{g}} = (\hat{g}_1, \dots, \hat{g}_H) = (\hat{p}_1^{1/2}, \dots, \hat{p}_H^{1/2})$ . Define

$$\begin{aligned} M &= [\boldsymbol{\vartheta}_{1,(m)}, \dots, \boldsymbol{\vartheta}_{H,(m)}]_{m \times H}, & G &= \text{diag}\{g_1, \dots, g_H\}, \\ F &= G \mathcal{J}_{\mathbf{g}}, & B_{(m)} &= MF, \\ \hat{M} &= [\hat{\boldsymbol{\vartheta}}_{1,(m)}, \dots, \hat{\boldsymbol{\vartheta}}_{H,(m)}]_{m \times H}, & \hat{G} &= \text{diag}\{\hat{g}_1, \dots, \hat{g}_H\}, \\ \hat{F} &= \hat{G} \mathcal{J}_{\hat{\mathbf{g}}}, & \hat{B}_{(m)} &= \hat{M} \hat{F}, \end{aligned}$$

where  $\boldsymbol{\vartheta}_{h,(m)}$  and  $\hat{\boldsymbol{\vartheta}}_{h,(m)}$  are defined in (2.9) and (2.11), respectively. Thus, the inverse-regression covariance matrices  $V_{(m)}$  and  $\hat{V}_{(m)}$  can be rewritten

as

$$(2.12) \quad V_{(m)} = B_{(m)} B_{(m)}^T, \quad \widehat{V}_{(m)} = \widehat{B}_{(m)} \widehat{B}_{(m)}^T.$$

**3. Deciding the dimension of EDR space.** As explained in previous sections, we are particularly interested in functional data or high-dimensional multivariate data. Existing methods for deciding the dimensionality of EDR space in the multivariate setting [Li (1991), Schott (1994)] are not directly applicable to the types of data that are focused on in this paper. Ferré and Yao (2003, 2005) used a graphical approach to determine the number of EDR directions for functional data but a formal statistical procedure has been lacking.

Our approach is generically described as follows. To decide the dimension of the EDR space, as in Li (1991), we will conduct sequential testing of  $H_0: K \leq K_0$  versus  $H_a: K > K_0$  for  $K_0 = 0, 1, 2, \dots$ ; we will stop at the first instance  $K_0 = \widehat{K}$  when the test fails to reject  $H_0$  and declare  $\widehat{K}$  as the true dimension. Below, we consider two types of tests in the sequential testing procedure motivated by (2.8). In Section 3.1, we assume that  $m$  is fixed, while in Section 3.2 we consider  $m$  in a wide range.

**3.1. Chi-squared tests based on a fixed  $m$ .** Fix an  $m$  and focus on the between-slice inverse covariance matrix  $V_{(m)}$ , which has dimension  $m \times m$ ; recall that it only makes sense to consider  $m$  such that  $m \leq n - 1$  and, if  $X$  is a  $p$ -dimensional vector,  $m \leq p$ . Define

$$K_{(m)} = \text{rank}(V_{(m)}).$$

Clearly,  $K_{(m)} \leq K$  for all  $m$ . It is desirable to pick an  $m$  such that  $K_{(m)} = K$ . Note that this condition means that the projections of all of the EDR directions onto the space spanned by the first  $m$  principal components are linearly independent, which is very different from saying that all of the EDR directions are completely in the span of the first  $m$  principle components; see the examples in Section 4. However, picking an  $m$  to guarantee  $K_{(m)} = K$  before analyzing the data is clearly not always possible. A practical approach is to simply pick an  $m$  such that the first  $m$  principal components explain a large proportion, say, 95%, of the total variation in the  $X_i$ 's. Such an approach will work for most real-world applications. Still, keeping  $m$  fixed has its limitations. We will address them in more detail in future sections.

In the following, let  $\lambda_j(M)$  denotes the  $j$ th largest eigenvalue of a nonnegative-definite square matrix  $M$ . Under  $H_0: K \leq K_0$ , we have  $\lambda_{K_0+1}(V_{(m)}) = \dots = \lambda_m(V_{(m)}) = 0$ . Consider the test statistic

$$(3.1) \quad \mathcal{T}_{K_0, (m)} = n \sum_{j=K_0+1}^m \lambda_j(\widehat{V}_{(m)}).$$



Since  $\widehat{V}_{(m)}$  estimates  $V_{(m)}$ , large values of  $\mathcal{T}_{K_0,(m)}$  will support the rejection of  $H_0$ . The following theorem provides the asymptotic distribution of  $\mathcal{T}_{K_0,(m)}$  under  $H_0$ . For the convenience of the proofs, we will assume below that the positive eigenvalues of  $\Gamma_X$  are all distinct.

The following addresses the case where  $X$  is a Gaussian process.

**THEOREM 3.1.** *Suppose that (C1) holds and  $X$  is a Gaussian process. Assume that  $K \leq K_0$ , and let  $H > K_0 + 1$  and  $m \geq K_0 + 1$ . Denote by  $\mathcal{X}$  a random variable having a  $\chi^2$  distribution with  $(m - K_0)(H - K_0 - 1)$  degrees of freedom.*

(i) *If  $K_{(m)} = K_0$ , then*

$$(3.2) \quad \mathcal{T}_{K_0,(m)} \xrightarrow{d} \mathcal{X} \quad \text{as } n \rightarrow \infty.$$

(ii) *If  $K_{(m)} < K_0$ , then  $\mathcal{T}_{K_0,(m)}$  is asymptotically stochastically bounded by  $\mathcal{X}$ ; namely,*

$$\limsup_{n \rightarrow \infty} \mathbb{P}(\mathcal{T}_{K_0,(m)} > x) \leq \mathbb{P}(\mathcal{X} > x) \quad \text{for all } x.$$

Theorem 3.1 suggests a  $\chi^2$  test for testing  $H_0: K \leq K_0$  versus  $H_a: K > K_0$ , which is an extension of a test in Li (1991) for multivariate data. Ideally, case (i) holds and the  $\chi^2$  test has the correct size asymptotically, as  $n \rightarrow \infty$ . For a variety of reasons case (ii) may be true, for which the  $\chi^2$  test will be conservative. This point will be illustrated graphically by a simulation example in Figure 1 in Section 4.

The proof of Theorem 3.1 is highly nontrivial, which goes considerably beyond the scope of the multivariate counterpart. A theoretical result that is needed to establish (ii) of Theorem 3.1 appears to be new and is stated here.

**PROPOSITION 3.2.** *Let  $Z$  be a  $p \times q$  random matrix and we write  $Z = [Z_1|Z_2]$  where  $Z_1$  and  $Z_2$  have sizes  $p \times r$  and  $p \times (q - r)$ , respectively, for some  $0 < r < \min(p, q)$ . Assume that  $Z_1$  and  $Z_2$  are independent, and  $Z_2$  contains i.i.d.  $\text{Normal}(0, 1)$  entries. Then  $\sum_{j=r+1}^p \lambda_j(ZZ^T)$  is stochastically bounded by  $\chi^2$  with  $(p - r)(q - r)$  degrees of freedom.*

The case where  $Z$  is a matrix of i.i.d.  $\text{Normal}(0, 1)$  entries can be viewed as the special case,  $r = 0$ , in Proposition 3.2. In that case, the bound is the exact distribution since  $\sum_{j=1}^p \lambda_j(ZZ^T)$  equals the sum of squares of all of the entries of  $Z$  and is therefore distributed as  $\chi^2$  with  $pq$  degrees of freedom.

Next, we address the scenario where  $X$  is elliptically contoured but not necessarily Gaussian. Let

$$(3.3) \quad \tau_h = \mathbb{E}(\Theta^2 | Y \in S_h), \quad h = 1, \dots, H.$$

If  $K_{(m)} = K_0$ , then it can be seen from the proofs in the [Appendix](#) that

$$(3.4) \quad \mathcal{T}_{K_0, (m)} \xrightarrow{d} \sum_{k=1}^{H-K_0-1} \delta_k \mathcal{X}_k \quad \text{as } n \rightarrow \infty,$$

where  $\mathcal{X}_k$ 's are distributed as i.i.d.  $\chi^2$  with  $m - K_0$  degrees of freedom, and  $\delta_1, \dots, \delta_{H-K_0-1}$  are the eigenvalues of  $\Lambda \Xi \Lambda$ , with  $\Xi = \mathcal{J}_{\mathbf{g}} \{I - B_{(m)}^T (B_{(m)} \times B_{(m)}^T)^- B_{(m)}\} \mathcal{J}_{\mathbf{g}}$  and  $\Lambda = \text{diag}(\tau_1^{1/2}, \dots, \tau_H^{1/2})$ . If  $X$  is Gaussian, then  $\tau_h$ 's and  $\delta_k$ 's are identically equal to 1. In general, the limiting null distribution in (3.4) depends on the unknown parameters  $\delta_k$ . Cook (1998) suggested carrying out this type of test by simulating the critical regions based on the estimated values of these parameters. Below, we introduce a different approach by adjusting the test statistic so that the limiting distribution is free of nuisance parameters.

Under  $H_0: K \leq K_0$ , let  $m > K_0$  and  $\widehat{\mathcal{P}}_2$  be the matrix whose columns are the eigenvectors that correspond to the  $m - K_0$  smallest eigenvalues of  $\widehat{V}_{(m)}$ . The definition (3.3) suggests (see proof of Theorem 3.3 in the [Appendix](#)) that  $\tau_h$  can be estimated by

$$(3.5) \quad \widehat{\tau}_h = \frac{1}{(m - K_0)n_h} \text{tr} \left\{ \widehat{\mathcal{P}}_2 \widehat{\mathcal{P}}_2^T \sum_{i=1}^n (\widehat{\boldsymbol{\eta}}_{i, (m)} - \widehat{\boldsymbol{\vartheta}}_{h, (m)}) \right. \\ \left. \times (\widehat{\boldsymbol{\eta}}_{i, (m)} - \widehat{\boldsymbol{\vartheta}}_{h, (m)})^T I(Y_i \in S_h) \right\}.$$

Put  $\Lambda = \text{diag}(\tau_1^{1/2}, \dots, \tau_H^{1/2})$ ,  $\widehat{\Lambda} = \text{diag}(\widehat{\tau}_1^{1/2}, \dots, \widehat{\tau}_H^{1/2})$ , and define

$$W_{(m)} = B_{(m)} \Lambda (\Lambda \mathcal{J}_{\mathbf{g}} \Lambda)^-, \quad \Sigma_{(m)} = W_{(m)} W_{(m)}^T, \\ \widehat{W}_{(m)} = \widehat{B}_{(m)} \widehat{\Lambda} (\widehat{\Lambda} \mathcal{J}_{\widehat{\mathbf{g}}} \widehat{\Lambda})^-, \quad \widehat{\Sigma}_{(m)} = \widehat{W}_{(m)} \widehat{W}_{(m)}^T,$$

where  $A^-$  denotes the Moore–Penrose generalized inverse of the matrix  $A$ . By Lemma 3 below,  $\Sigma_{(m)}$  has the same null space as  $V_{(m)}$ . Thus, under  $H_0: K \leq K_0$ , we have  $\lambda_{K_0+1}(\Sigma_{(m)}) = \dots = \lambda_m(\Sigma_{(m)}) = 0$ . We therefore propose the test statistic

$$\mathcal{T}_{K_0, (m)}^* = n \sum_{j=K_0+1}^m \lambda_j(\widehat{\Sigma}_{(m)}),$$

where, again, large values of  $\mathcal{T}_{K_0, (m)}^*$  support the rejection of  $H_0$ . The following result extends (i) of Theorem 3.1 from the case where  $X(t)$  is Gaussian to a general elliptically contoured process. While we conjecture that (ii) of Theorem 3.1 can be similarly extended, we have not been able to prove it.

**THEOREM 3.3.** *Suppose that (C1) and (C3) hold. Assume that the true dimension  $K \leq K_0$  and let  $H > K_0 + 1$  and  $m \geq K_0 + 1$ . If  $K_{(m)} = K_0$  then  $\mathcal{T}_{K_0, (m)}^* \xrightarrow{d} \chi_{(m-K_0)(H-K_0-1)}^2$  as  $n \rightarrow \infty$ .*

The test of  $H_0: K \leq K_0$  based on  $\mathcal{T}_{K_0, (m)}$  and  $\mathcal{T}_{K_0, (m)}^*$  and the asymptotic null distribution,  $\chi_{(m-K_0)(H-K_0-1)}^2$ , will be referred to as the  $\chi^2$  test and the adjusted  $\chi^2$  test, respectively.

**3.2. Adaptive Neyman tests.** So far we considered tests based on a fixed  $m$ . In most situations, in practice, choosing the smallest  $m$  for which the first  $m$  sample principal components explain most of the variations should work fairly well for determining  $K$ . However, for functional or high-dimensional multivariate data, one cannot theoretically rule out the possibility that the EDR directions can only be detected by examining the information contained in higher-order principal components.

A careful inspection reveals two different issues here. The first is the question that if we have an unusual model in which some EDR directions depend only on high-order principal components that the data have little power to discern, can any approach be effective in detecting the presence of those directions? The answer is “not likely” since, intuitively, we can detect the presence of those EDR directions no better than we can the principle components that comprise the directions. This is due more to the nature of high or infinite-dimensional data than the limitation of any methodology. However, keep in mind that principal components are not ordinary covariates, but are mathematical constructs which not only depend on the covariance function of  $X$  but also the choice of inner product of  $\mathcal{H}$ . Thus, one can argue that having an EDR direction that is orthogonal to a large number of low-order principal components of the predictor is itself a rather artificial scenario and is not likely to be the case in practice.

Let us now turn to the second issue. Assume that all of the EDR directions do contain low-order principal components which can be estimated well from data. For example, suppose each EDR direction is not in the orthogonal complement of the space spanned by the first three principal components and so the procedures described in Section 3.1 will in principle work if we let  $m = 3$ . However, since that knowledge is not available when we conduct data analysis, to be sure perhaps we might consider picking a much larger truncation point, say,  $m = 30$ . The problem with this approach is that, when the sample size is fixed, the power of the tests will decrease with  $m$ . Intuitively, when  $m$  is large the information contained in the individual components of  $\hat{\vartheta}_{h, (m)} = (\hat{\vartheta}_{1, h}, \dots, \hat{\vartheta}_{m, h})^T$  becomes diluted. We will illustrate this point numerically in Section 4.1. This is strikingly similar to the situation of testing whether the mean of a high-dimensional normal random vector is nonzero

described at the beginning of Section 2 of Fan and Lin (1998), where the power of the Neyman test (likelihood-ratio test) was shown to converge to the size of the test as the number of dimension increases. Essentially, the problem that they describe is caused by the fact that the Neyman test has a rejection region that is symmetric in all components of the vector, which is designed to treat all possible alternatives uniformly. Fan and Lin (1998) argued that the alternatives that are of the most importance in practice are usually those in which the leading components of the Gaussian mean vector are nonzero, and they modified the Neyman test accordingly such that the test will have satisfactory powers for those alternatives.

We now introduce a test inspired by Fan and Lin (1998) that avoids having to pick a specific  $m$ . To test  $H_0: K \leq K_0$  against  $H_a: K > K_0$ , we compute  $\mathcal{T}_{K_0,(m)}$  for all of  $m = K_0 + 1, \dots, N$ , for some “large”  $N$ ; we then take the maximum of the standardized versions of these test statistics, and the null hypothesis will be rejected for large values of the maximum. To facilitate this approach, we present the following result that is a deeper version of Theorem 3.1 and shows that the test statistics  $\mathcal{T}_{K_0,(m)}$  has a “partial sum” structure in  $m$  as the sample size tends to  $\infty$ .

**THEOREM 3.4.** *Suppose that (C1) holds and  $X$  is a Gaussian process. Assume that  $K \leq K_0$  and let  $H > K_0 + 1$ . Let  $\chi_i^2, i \geq 1$ , be i.i.d.  $\chi^2$  random variables with  $H - K_0 - 1$  degrees of freedom and define  $\mathcal{X}_{(m)} = \sum_{i=1}^{m-K_0} \chi_i^2, m \geq K_0 + 1$ . Then, for all positive integers  $N > K_0$ , the collection of test statistics  $\mathcal{T}_{K_0,(m)}, m = K_0 + 1, \dots, N$ , are jointly stochastically bounded by  $\mathcal{X}_{(m)}, m = K_0 + 1, \dots, N$ , as the sample size  $n$  tends to  $\infty$ .*

In view of Theorem 3.4, we propose the following. To test  $H_0: K \leq K_0$  versus  $H_a: K > K_0$ , define

$$\mathcal{U}_{K_0,N} := \max_{K_0+1 \leq m \leq N} \frac{\mathcal{T}_{K_0,(m)} - (m - K_0)(H - K_0 - 1)}{\sqrt{2(m - K_0)(H - K_0 - 1)}},$$

and we reject  $H_0$  at level  $\alpha$  if  $\mathcal{U}_{K_0,N} > u_\alpha$  where  $u_\alpha$  is the  $1 - \alpha$  quantile of

$$\mathcal{B}_{K_0,N} := \max_{K_0+1 \leq m \leq N} \frac{\mathcal{X}_{(m)} - (m - K_0)(H - K_0 - 1)}{\sqrt{2(m - K_0)(H - K_0 - 1)}}.$$

The resulting test resembles asymptotically the aforementioned test in Fan and Lin (1998) which was referred to as an adaptive Neyman test. For convenience, we will also refer to our test as adaptive Neyman test, although the contexts of the two problems are completely unrelated.

Suppose that  $H_0$  holds and  $m_{K_0}$  is the smallest  $m$  such that  $K_{(m)} = K_0$ . Then, by Theorem 3.1,  $\mathcal{U}_{K_0,N} - \mathcal{U}_{K_0,m_{K_0}-1} \xrightarrow{d} \mathcal{B}_{K_0,N} - \mathcal{B}_{K_0,m_{K_0}-1}$ . Thus,  $\mathcal{B}_{K_0,N}$  is, intuitively, a tight asymptotic stochastic bound for  $\mathcal{U}_{K_0,N}$ .

Simulation results show that the maximum in the definition of  $\mathcal{U}_{K_0, N}$  is, with high probability, attained at relatively small  $m$ 's. Thus, the test is quite robust with respect to the choice of  $N$ . In practice, one can pick  $N$  so that there is virtually just noise beyond the  $N$ th sample principal component. Numerically, the performance of the adaptive Neyman test matches those of the  $\chi^2$  tests in which  $m$  is chosen correctly, but does not have the weakness of possibly under-selecting  $m$ .

### 3.3. Discussion.

- (a) Our procedures apply to both finite-dimensional and infinite-dimensional data, and, in particular, are useful for treating high-dimensional multivariate data. In that case, Li's  $\chi^2$  test suffers from the problem of diminishing power as does the test developed in Schott (1994); see, for example, Table 4 in Schott (1994). Our procedures can potentially be a viable solution in overcoming the power loss problem in that situation. The inclusion of measurement error in  $X$  provides additional flexibility in modeling multivariate data. Note that the formulation of Theorem 2.1 can be extended to accommodate measurement error: if  $X = X_1 + X_2$  where  $X_1$  is the true covariate with mean  $\mu$  and covariance matrix  $\Gamma_{X_1}$  and  $X_2$  is independent measurement error with mean zero, then  $\mathbb{E}(X|Y) = \mathbb{E}(X_1|Y)$  and so  $\text{Im}(\Gamma_{X|Y}) \subset \text{span}(\Gamma_{X_1}\beta_1, \dots, \Gamma_{X_1}\beta_K)$ . Thus, one might speculate that our procedures continue to work in that case, and this is borne out by simulations presented in Section 4.3. Detailed theoretical investigation of this is a topic of future work, but preliminary indications are that the extension of Theorem 2.1 is valid at least under the additional assumption that the components of  $X_2$  are i.i.d. with finite variance.
- (b) Choice of slices: in the SIR literature, the prevailing view is that the choice of slices is of secondary importance. In our simulation studies, we used contiguous slices containing roughly the same number of  $Y_i$ 's, where the number of  $Y_i$ 's per slice that we experimented with ranged from 25 to 65. Within this range, we found that the number of data per slice indeed had a negligible effect on the estimation of  $K$ .
- (c) Choice of  $\alpha$ : if  $\alpha$  is fixed and  $m$  and  $N$  are chosen sensibly in the  $\chi^2$  tests and the adaptive Neyman test, respectively, then the asymptotic results show that the probability of correct identification of  $K$  tends to  $1 - \alpha$  as  $n$  tends to  $\infty$ . In real-data applications, the optimal choice of  $\alpha$  depends on a number of factors including the sample size and the true model. In our simulation studies, presented in Section 4,  $\alpha = 0.05$  worked well for all of our settings.
- (d) Limitations of SIR: the failure of SIR in estimating the EDR space in situations where  $Y$  depends on  $X$  in a symmetric manner is well

documented. While exact symmetry is not a highly probable scenario in practice, it does represent an imperfection of SIR which has been addressed by a number of other methods including SAVE in Cook and Weisberg (1991), MAVE in Xia et al. (2002) and integral transform methods in Zhu and Zeng (2006, 2008). The estimation of  $K$  based on those approaches will be a topic of future research.

#### 4. Simulation studies.

4.1. *Simulation 1: Sizes and power of the tests.* In this study, we consider functional data generated from the process

$$X(t) = \sum_{k=1}^{\infty} \omega_{2k-1}^{1/2} \eta_{2k-1} \sqrt{2} \cos(2k\pi t) + \sum_{k=1}^{\infty} \omega_{2k}^{1/2} \eta_{2k} \sqrt{2} \sin(2k\pi t),$$

where  $\omega_k = 20(k + 1.5)^{-3}$ . Thus, the principal components are the sine and cosine curves in the sum. We will consider the cases where the  $\eta_k$ 's follow  $\text{Normal}(0, 1)$  and centered multivariate  $t$  distribution with  $\nu = 5$  degrees of freedom, with the latter representing the situation where  $X$  is an elliptically contoured process. Note that the centered multivariate  $t$  distribution with  $\nu$  degrees of freedom can be represented by

$$\mathbf{t} = Z / \sqrt{\tau / (\nu - 2)} \sim t_{\nu} \quad \text{where } Z \sim N(0, I) \text{ and } \tau \sim \chi_{\nu}^2 \text{ are independent.}$$

To simulate  $\{\eta_1, \eta_2, \dots\}$  in that case, we first simulate  $z_1, z_2, \dots \sim \text{i.i.d. Normal}(0, 1)$ ,  $\tau \sim \chi_{\nu}^2$ , and then put  $\eta_k = z_k / \sqrt{\tau / (\nu - 2)}$ . By this construction, any finite collection of the  $\eta_k$ 's follows a multivariate  $t$  distribution, where the  $\eta_k$ 's are mutually uncorrelated but not independent.

Let the EDR space be generated by the functions

$$\begin{aligned} \beta_1(t) &= 0.9\sqrt{2} \cos(2\pi t) + 1.2\sqrt{2} \cos(4\pi t) \\ &\quad - 0.5\sqrt{2} \cos(8\pi t) + \sum_{k>4} \frac{\sqrt{2}}{(2k-1)^3} \cos(2k\pi t), \\ \beta_2(t) &= -0.4\sqrt{2} \sin(2\pi t) + 1.5\sqrt{2} \sin(4\pi t) - 0.3\sqrt{2} \sin(6\pi t) \\ &\quad + 0.2\sqrt{2} \sin(8\pi t) + \sum_{k>4} \frac{(-1)^k \sqrt{2}}{(2k)^3} \sin(2k\pi t), \\ \beta_3(t) &= \sqrt{2} \cos(2\pi t) + \sqrt{2} \sin(4\pi t) + 0.5\sqrt{2} \cos(6\pi t) + 0.5\sqrt{2} \sin(8\pi t) \\ &\quad + \sum_{k\geq 3} \frac{\sqrt{2}}{(4k-3)^3} \cos\{2(2k-1)\pi t\} + \sum_{k\geq 3} \frac{\sqrt{2}}{(4k)^3} \sin(4k\pi t). \end{aligned} \tag{4.1}$$

Consider the models

$$\text{Model 1: } Y = 1 + 2 \sin(\langle \beta_1, X \rangle) + \varepsilon,$$

$$\text{Model 2: } Y = \langle \beta_1, X \rangle \times (2\langle \beta_2, X \rangle + 1) + \varepsilon,$$

$$\text{Model 3: } Y = 5\langle \beta_1, X \rangle \times (2\langle \beta_2, X \rangle + 1) / (1 + \langle \beta_3, X \rangle^2) + \varepsilon,$$

where  $\varepsilon \sim \text{Normal}(0, 0.5^2)$ . The EDR spaces of the three models have dimensions  $K = 1, 2$  and  $3$ , respectively. Also note that  $K_{(m)} = K$  if  $m \geq 1, 2$  and  $3$ , respectively, for the three models.

In each of 1000 simulation runs, data were simulated from models 1–3 for the two distributional scenarios that  $X$  is distributed as Gaussian and  $t$  with two sample sizes  $n = 200$  and  $500$ . To mimic real applications, we assumed that each curve  $X_i$  is observed at 501 equally-spaced points. We then registered the curves using 100 Fourier basis functions. A functional principal component analysis was carried out using the package `fda` in R contributed by Jim Ramsay.

To decide the dimension of the EDR space, we compared the two proposed  $\chi^2$  tests and the adaptive Neyman test. For the  $\chi^2$  tests, we let  $m = 5, 7$  and  $30$ , where the first 5, 7 and 30 principal components of  $X$ , respectively, account for 91%, 95% and 99.59% of the total variation. We present the results for  $m = 30$  as an extreme case to illustrate the point that using a large number of principal components will cause the tests to have lower powers. For the adaptive Neyman test, we took  $N = K_0 + 30$  and simulated the critical values for  $\mathcal{U}_{K_0, N}$  based on the description following Theorem 3.4. We only report the results based on  $H = 8$  slices, but the choice was not crucial. The nominal size of the tests was set to be  $\alpha = 0.05$ .

The simulation results are briefly discussed below. Table 1 gives the empirical frequencies of rejecting  $H_0: K \leq 1$ . Since the dimension of EDR space under model 1 is equal to 1, the results in the column under model 1 give the empirical sizes of the tests. Models 2 and 3 represent two cases under the alternative hypothesis, therefore the results in those columns give the power of the tests. As can be seen, when  $m$  is 5 or 7, the two  $\chi^2$  tests have sizes close to the nominal size and have high powers. However, for the case  $m = 30$  and  $n = 200$ , the tests performed significantly worse in those metrics. On the other hand the adaptive Neyman test performs very stably, with powers comparable to those of  $\chi^2$  tests with a well chosen  $m$ . It is also worth noting that when  $X$  has the  $t$  distribution, the  $\chi^2$  test performs comparably to, sometimes better than, the adjusted  $\chi^2$  test, showing that the  $\chi^2$  test is quite robust against departure from normality.

Table 2 shows that the empirical frequencies of finding the true dimensions for different situations. As can be expected, estimating the true dimension becomes more challenging as the model becomes more complicated. For example, the probabilities of finding the true dimension for model 3 are

TABLE 1

*Empirical frequencies of rejecting the hypothesis  $H_0: K \leq 1$ . The results are based on 1000 simulations for each of the three models, two distributions of process  $X(t)$ , and two sample sizes. The  $\chi^2$  test and the adjusted  $\chi^2$  are applied with fixed  $m$  values, and the adaptive Neyman test is applied with  $N = K_0 + 30$*

|                        |                            | Model 1 |       | Model 2 |       | Model 3 |       |
|------------------------|----------------------------|---------|-------|---------|-------|---------|-------|
| Distribution of $\eta$ |                            | Normal  | $t$   | Normal  | $t$   | Normal  | $t$   |
| $n = 200$              | $\chi^2$ test ( $m = 5$ )  | 0.040   | 0.046 | 0.883   | 0.567 | 0.995   | 0.949 |
|                        | Adj. $\chi^2$ ( $m = 5$ )  | 0.070   | 0.082 | 0.894   | 0.584 | 0.997   | 0.954 |
|                        | $\chi^2$ test ( $m = 7$ )  | 0.045   | 0.039 | 0.827   | 0.496 | 0.982   | 0.910 |
|                        | Adj. $\chi^2$ ( $m = 7$ )  | 0.071   | 0.083 | 0.868   | 0.537 | 0.989   | 0.924 |
|                        | $\chi^2$ test ( $m = 30$ ) | 0.034   | 0.020 | 0.320   | 0.111 | 0.677   | 0.493 |
|                        | Adj. $\chi^2$ ( $m = 30$ ) | 0.105   | 0.072 | 0.484   | 0.299 | 0.798   | 0.668 |
|                        | Adaptive Neyman            | 0.044   | 0.045 | 0.860   | 0.545 | 0.993   | 0.936 |
| $n = 500$              | $\chi^2$ test ( $m = 5$ )  | 0.052   | 0.043 | 1.000   | 0.977 | 1.000   | 1.000 |
|                        | Adj. $\chi^2$ ( $m = 5$ )  | 0.069   | 0.056 | 1.000   | 0.969 | 1.000   | 1.000 |
|                        | $\chi^2$ test ( $m = 7$ )  | 0.048   | 0.033 | 1.000   | 0.958 | 1.000   | 0.999 |
|                        | Adj. $\chi^2$ ( $m = 7$ )  | 0.056   | 0.049 | 1.000   | 0.949 | 1.000   | 0.999 |
|                        | $\chi^2$ test ( $m = 30$ ) | 0.059   | 0.025 | 0.958   | 0.584 | 0.999   | 0.990 |
|                        | Adj. $\chi^2$ ( $m = 30$ ) | 0.085   | 0.054 | 0.972   | 0.666 | 0.999   | 0.991 |
|                        | Adaptive Neyman            | 0.052   | 0.040 | 1.000   | 0.963 | 1.000   | 0.999 |

much smaller than those for model 2. Our simulation results also show that, for a range of small values of  $m$ , the two  $\chi^2$  procedures perform very well especially if  $n = 500$ , where for brevity those results are represented by  $m = 5$  and 7 in Table 2. However, when  $m = 30$ , the probabilities of finding the true dimension become smaller for those procedures, which is especially true for models 2 and 3. This is another illustration that using a large number of principal components will lead to a loss of power for the underlying  $\chi^2$  tests. Again, the adaptive Neyman procedure performs comparably to the two  $\chi^2$  procedures with a well-chosen  $m$ .

*4.2. Simulation 2: Sensitivity to the truncation point  $m$ .* In the examples in Section 4.1, the three EDR directions are linearly independent when projected onto the three leading principal components. Hence, the two  $\chi^2$  procedures are expected to work so long as  $m \geq 3$ . For situations where one or more EDR directions only depend on high-order principal components, the choice of  $m$  in the  $\chi^2$  procedure is crucial and the adaptive Neyman procedure has a clear advantage.

To illustrate this, we consider a new model

$$\text{Model 4: } Y = \langle \beta_1, X \rangle \times (2\langle \beta_4, X \rangle + 1) + \varepsilon,$$



TABLE 2

*Empirical frequencies of finding the true dimension of the EDR space. The results are based on 1000 simulations for each of the three models, two distributions of process  $X(t)$ , and two sample sizes. The  $\chi^2$  test and the adjusted  $\chi^2$  are applied with fixed  $m$  values, and the adaptive Neyman test is applied with  $N = K_0 + 30$*

|                        |                            | Model 1 |       | Model 2 |       | Model 3 |       |
|------------------------|----------------------------|---------|-------|---------|-------|---------|-------|
| Distribution of $\eta$ |                            | Normal  | $t$   | Normal  | $t$   | Normal  | $t$   |
| $n = 200$              | $\chi^2$ test ( $m = 5$ )  | 0.960   | 0.954 | 0.859   | 0.562 | 0.322   | 0.165 |
|                        | Adj. $\chi^2$ ( $m = 5$ )  | 0.930   | 0.918 | 0.846   | 0.556 | 0.393   | 0.224 |
|                        | $\chi^2$ test ( $m = 7$ )  | 0.955   | 0.961 | 0.809   | 0.488 | 0.272   | 0.116 |
|                        | Adj. $\chi^2$ ( $m = 7$ )  | 0.929   | 0.917 | 0.832   | 0.505 | 0.313   | 0.167 |
|                        | $\chi^2$ test ( $m = 30$ ) | 0.966   | 0.971 | 0.309   | 0.111 | 0.057   | 0.026 |
|                        | Adj. $\chi^2$ ( $m = 30$ ) | 0.895   | 0.924 | 0.462   | 0.277 | 0.119   | 0.079 |
|                        | Adaptive Neyman            | 0.956   | 0.955 | 0.843   | 0.542 | 0.337   | 0.158 |
| $n = 500$              | $\chi^2$ test ( $m = 5$ )  | 0.948   | 0.957 | 0.955   | 0.958 | 0.842   | 0.629 |
|                        | Adj. $\chi^2$ ( $m = 5$ )  | 0.931   | 0.944 | 0.948   | 0.910 | 0.849   | 0.648 |
|                        | $\chi^2$ test ( $m = 7$ )  | 0.952   | 0.967 | 0.959   | 0.948 | 0.739   | 0.489 |
|                        | Adj. $\chi^2$ ( $m = 7$ )  | 0.944   | 0.951 | 0.949   | 0.898 | 0.754   | 0.551 |
|                        | $\chi^2$ test ( $m = 30$ ) | 0.941   | 0.975 | 0.913   | 0.582 | 0.279   | 0.101 |
|                        | Adj. $\chi^2$ ( $m = 30$ ) | 0.915   | 0.946 | 0.910   | 0.625 | 0.308   | 0.203 |
|                        | Adaptive Neyman            | 0.948   | 0.960 | 0.967   | 0.952 | 0.843   | 0.613 |

where  $X$  is a Gaussian process whose distribution is as described in Section 4.1,  $\beta_1$  is as in (4.1), but  $\beta_4$  is given by

$$\begin{aligned} \beta_4(t) = & 0.45\sqrt{2}\cos(2\pi t) + 0.6\sqrt{2}\cos(4\pi t) - 3\sqrt{2}\sin(6\pi t) \\ & + 1.2\sqrt{2}\sin(8\pi t) + \sum_{k>4} \frac{(-1)^k\sqrt{2}}{(2k)^3} \sin(2k\pi t). \end{aligned}$$

In this model, the dimension of the EDR space is 2, but the projections of  $\beta_1$  and  $\beta_4$  onto the first five principal components are linearly dependent; indeed,  $K_{(m)} = 1, m \leq 5$ , and  $K_{(m)} = 2, m \geq 6$ . As shown in Table 3, the two  $\chi^2$  procedures with  $m = 5$  both failed to find the true dimension, even when  $n = 500$ . On the other hand, when  $n = 500$  and  $m = 7$ , the  $\chi^2$  procedures worked very well. With  $m = 30$ , both  $\chi^2$  tests again have considerably lower powers, which leads to smaller probabilities of correct identification. As in the previous models, the adaptive Neyman procedure has comparable performance to the best  $\chi^2$  procedures.

Finally, we use model 4 to illustrate the null distribution of  $\mathcal{T}_{K_0,(m)}$  and  $\mathcal{T}_{K_0,(m)}^*$  when  $K_{(m)} < K_0$ . Consider  $K_0 = 2$ ; for each  $m = 4, 5, \dots$ , compute the expected values of  $\mathcal{T}_{K_0,(m)}$  and  $\mathcal{T}_{K_0,(m)}^*$  by simulations and compare

TABLE 3  
Empirical frequencies of finding the correct model in model  
4

|                            | $n = 200$ | $n = 500$ |
|----------------------------|-----------|-----------|
| $\chi^2$ test ( $m = 5$ )  | 0.040     | 0.047     |
| Adj. $\chi^2$ ( $m = 5$ )  | 0.068     | 0.068     |
| $\chi^2$ test ( $m = 7$ )  | 0.358     | 0.913     |
| Adj. $\chi^2$ ( $m = 7$ )  | 0.410     | 0.899     |
| $\chi^2$ test ( $m = 30$ ) | 0.085     | 0.566     |
| Adj. $\chi^2$ ( $m = 30$ ) | 0.170     | 0.616     |
| Adaptive Neyman            | 0.229     | 0.885     |

the expectations with theoretical expectations  $(m - K_0)(H - K_0 - 1)$ . The results are described in Figure 1, in which the grey rectangles mark the means of  $\mathcal{T}_{K_0, (m)}$ , the black circles mark the means of  $\mathcal{T}_{K_0, (m)}^*$ , and straight line represents  $(m - K_0)(H - K_0 - 1)$ . The case  $m \geq 6$  correspond to (i) of Theorem 3.1, for which  $\chi_{(m-K_0)(H-K_0-1)}^2$  is the asymptotic distribution of the test statistics; the cases  $m = 4$  and 5 correspond to (ii) of Theorem 3.1, for which  $\chi_{(m-K_0)(H-K_0-1)}^2$  is only a stochastic bound of the test statistics. Both of these points are clearly reflected in Figure 1.

4.3. *Simulation 3: Multivariate data with measurement errors.* In this subsection, we present a simulation study for high-dimensional multivariate data; in particular, we use the study to support claims made in (a) of

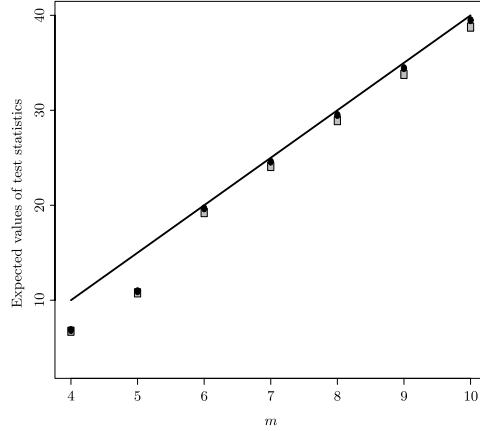


FIG. 1. The expected values of the test statistics  $\mathcal{T}_{K_0, (m)}$  and  $\mathcal{T}_{K_0, (m)}^*$  plotted as a function of  $m$ ; the solid line describes the theoretical expected values, the rectangles are the means of  $\mathcal{T}_{K_0, (m)}$  and the circles are the means of  $\mathcal{T}_{K_0, (m)}^*$ .

Section 3.3. Assume that  $X$  is multivariate, and, for clarity, denote  $X$  by  $\mathbf{X}$  here. The simulated model is described as follows. We first generate a  $p$ -dimensional variable  $\mathbf{X} = (\mathbf{X}_1^T, \mathbf{X}_2^T)^T$ , where  $p$  is “large,” with  $\mathbf{X}_1$  denoting a 10-dimensional random vector while  $\mathbf{X}_2 = \mathbf{0}_{p-10}$ , so that  $\mathbf{X}_1$  contains the real signal in  $\mathbf{X}$ . Suppose that  $\mathbf{X}_1$  has a low-dimensional representation

$$\mathbf{X}_1 = \sum_{k=1}^5 \xi_k \psi_k,$$

where the  $\psi_k$ ’s are orthonormal vectors,  $\xi_k \sim \text{Normal}(0, \omega_k)$  are independent, and  $(\omega_1, \dots, \omega_5) = (3, 2.8, 2.6, 2.4, 2.2)$ ; the  $\psi_k$ ’s are randomly generated, but are fixed throughout the simulation study. Furthermore, instead of observing the true  $\mathbf{X}$ , assume that we observe an error-prone surrogate,

$$\mathbf{W} = \mathbf{X} + \mathbf{U},$$

where  $\mathbf{U} \sim \text{Normal}(\mathbf{0}, \mathbf{I}_p)$  is measurement error. Thus, the eigenvalues of the covariance of  $\mathbf{W}$  are bounded below by 1. Note that this is a simpler measurement-error model than the one considered in Carroll and Li (1992), but realistically portrays certain crucial aspects of high-dimensional data encountered in practice; for example, in a typical fMRI study the total number of brain voxels captured by the image is huge but often only a relatively small portion of the voxels are active for the task being studied, while background noise is ubiquitous.

Let  $X_1, \dots, X_p$  be the components of  $\mathbf{X}$ . Consider the model

$$\text{Model 5: } Y = (X_1 + X_2)/(X_2 + X_3 + X_4 + X_5 + 1.5)^2 + \varepsilon,$$

where  $\varepsilon \sim \text{Normal}(0, 0.5^2)$ . Thus,  $Y$  only depends on  $\mathbf{X}_1$ . Below we compare the  $\chi^2$  procedure in Li (1991) and the adaptive Neyman procedure using  $\mathbf{W}$  as the observed covariate.

We conducted simulations for  $n = 200, 500$  and  $p = 15, 20, 40, 100$ . For each setting, we repeat the simulation for 1000 times and used Li’s procedure and the adaptive Neyman procedure in deciding the number of EDR directions. For both procedures, the nominal size  $\alpha = 0.05$  was used. In Table 4, we

TABLE 4  
Empirical frequencies of finding the correct model in model 5

|           |                    | $p = 15$ | $p = 20$ | $p = 40$ | $p = 100$ |
|-----------|--------------------|----------|----------|----------|-----------|
| $n = 200$ | Li’s $\chi^2$ test | 0.328    | 0.258    | 0.123    | 0.007     |
|           | Adaptive Neyman    | 0.588    | 0.596    | 0.562    | 0.528     |
| $n = 500$ | Li’s $\chi^2$ test | 0.898    | 0.833    | 0.612    | 0.276     |
|           | Adaptive Neyman    | 0.955    | 0.956    | 0.953    | 0.960     |

TABLE 5  
*Empirical sizes of Li's  $\chi^2$  test and the adaptive Neyman test for  $H_0: K \leq 2$  under model 5*

|           |                    | $p = 15$ | $p = 20$ | $p = 40$ | $p = 100$ |
|-----------|--------------------|----------|----------|----------|-----------|
| $n = 200$ | Li's $\chi^2$ test | 0.015    | 0.012    | 0.002    | 0.001     |
|           | Adaptive Neyman    | 0.016    | 0.013    | 0.009    | 0.010     |
| $n = 500$ | Li's $\chi^2$ test | 0.044    | 0.042    | 0.025    | 0.013     |
|           | Adaptive Neyman    | 0.037    | 0.035    | 0.035    | 0.030     |

summarize the empirical frequencies of finding the correct dimension. As can be seen, while the performance of the adaptive Neyman procedure is quite stable for different  $p$ 's, the performance of Li's procedure deteriorates as  $p$  increases. In Table 5, we also present the true sizes, obtained by simulations, of the two tests for  $H_0: K \leq 2$ . In all cases both tests have sizes under 0.05. The sizes of both tests are closer to the nominal size when  $n = 500$  than when  $n = 200$ . With a fixed  $n$ , the sizes of Li's test decrease quickly as  $p$  increases, reflecting the conservative nature of the test for large  $p$ , while those for the adaptive Neyman test remain relatively stable.

**5. Data analysis.** In this section, we consider the Tecator data [Thodberg (1996)], which can be downloaded at <http://lib.stat.cmu.edu/datasets/tecator>. The data were previously analyzed in a number of papers including Ferré and Yao (2005), Amato, Antoniadis and De Feis (2006) and Hsing and Ren (2009). The data contains measurements obtained by analyzing 215 meat samples, where for each sample a 100-channel, near-infrared spectrum was obtained by a spectrometer, and the water, fat and protein contents were also directly measured. The spectral data can be naturally considered as functional data, and we are interested in building a regression model to predict the fat content from the spectrum. Following the convention in the literature, we applied a logistic transformation to the percentage of fat content,  $U$ , by letting  $Y = \log_{10}\{U/(1 - U)\}$ .

In applying functional SIR, both Ferré and Yao (2005) and Amato, Antoniadis and De Feis (2006) used graphical tools to select the number EDR directions, where the numbers of directions selected were 10 and 8, respectively. On the other hand, using only two EDR directions, Amato, Antoniadis and De Feis (2006) applied MAVE to achieve a prediction error comparable to what can be achieved by SIR using 8 directions. These conclusions were somewhat inconsistent.

Based on the instructions given by the Tecator website, we used the first 172 samples for training and the last 43 for testing. Following Amato, Antoniadis and De Feis (2006), we focused on the most informative part of the spectra, with wavelengths ranging from 902 to 1028 nm. The curves are

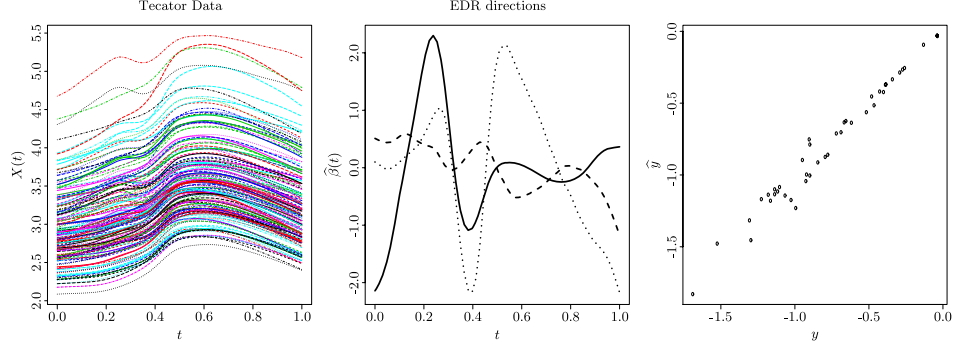


FIG. 2. *Tecator spectrum data: the first plot shows the Tecator spectrum data in the training set, the second plot show the estimated EDR directions and the last plot is the predicted vs. true fat contents for the test data set.*

rescaled onto the interval  $[0, 1]$ . The first plot in Figure 2 shows those spectra in the training set.

We first fitted B-splines to the discrete data, and then applied our sequential-testing procedures. With  $\alpha = 0.05$ , the adaptive Neyman procedure concluded that  $K = 3$ . To see how well a three-dimensional model works, the model  $Y = f(\langle \beta_1, X \rangle, \langle \beta_2, X \rangle, \langle \beta_3, X \rangle) + \varepsilon$  was entertained. The EDR directions were estimated by the regularized approach, RSIR, introduced by Zhong et al. (2005); the estimated EDR directions are presented in the center plot in Figure 2. Finally, the link function  $f$  was estimated by smoothing spline ANOVA [Gu (2002)] with interaction terms; the estimated model was then applied to test data to predict fat content. The root mean prediction error was 0.062 which is comparable to what was obtained by MAVE in Amato, Antoniadis and De Feis (2006). The plot of the predicted versus the true fat contents for test data is also given in Figure 2.

Our result is in agreement with what was obtained using MAVE in Amato, Antoniadis and De Feis (2006) in that a low-dimensional model is appropriate for this data set.

## APPENDIX: PROOFS

In the following, the notation “ $\star$ ” refers to symbolic matrix multiplication; for instance, if  $f_1, \dots, f_k$  are mathematical objects (functions, matrices, etc.) and  $\mathbf{c} = (c_1, \dots, c_k)^T$  is a vector for which the operation  $\sum_{i=1}^k c_i f_i$  is defined, we will denote the sum by  $(f_1, \dots, f_k) \star \mathbf{c}$ ; also if  $C$  is a matrix containing columns  $\mathbf{c}_1, \dots, \mathbf{c}_\ell$ , the notation  $(f_1, \dots, f_k) \star C$  refers the array  $[(f_1, \dots, f_k) \star \mathbf{c}_1, \dots, (f_1, \dots, f_k) \star \mathbf{c}_\ell]$ .

The notation is defined in Section 2 and will be used extensively below without further mention.

**A.1. Proof of Theorem 3.1.** Recall from (2.12) that  $V_{(m)} = B_{(m)}B_{(m)}^T$  and  $\hat{V}_{(m)} = \hat{B}_{(m)}\hat{B}_{(m)}^T$ . Thus, to study the eigenvalues of  $V_{(m)}$  and  $\hat{V}_{(m)}$ , we can equivalently study the singular values of  $B_{(m)}$  and  $\hat{B}_{(m)}$ , respectively. Recall that  $K_{(m)} = \text{Rank}(B_{(m)})$ . Under the hypothesis  $K \leq K_0$ , we also have  $K_{(m)} \leq K_0$ . Suppose  $B_{(m)}$  has the following singular-value decomposition:

$$B_{(m)} = \mathcal{P} \begin{pmatrix} D & 0 \\ 0 & 0 \end{pmatrix} \mathcal{Q}^T,$$

where  $D := \text{diag}(\lambda_1^{1/2}(V_{(m)}), \dots, \lambda_{K_{(m)}}^{1/2}(V_{(m)}))$  contains the nonzero singular values of  $B_{(m)}$ , and  $\mathcal{P}$  and  $\mathcal{Q}$  are orthonormal matrices of dimensions  $m \times m$  and  $H \times H$ , respectively, which contain the singular vectors of  $B_{(m)}$ . Note that, for brevity of notation, we leave out  $m$  in  $\mathcal{P}, \mathcal{Q}$  and  $n$  in  $\hat{B}_{(m)}, \hat{V}_{(m)}$  in this proof.

Partition  $\mathcal{P}$  and  $\mathcal{Q}$  as  $\mathcal{P} = [\mathcal{P}_1 | \mathcal{P}_2]$ ,  $\mathcal{Q} = [\mathcal{Q}_1 | \mathcal{Q}_2]$  where  $\mathcal{P}_1$  and  $\mathcal{Q}_1$  both have  $K_{(m)}$  columns, and  $\mathcal{P}_2$  and  $\mathcal{Q}_2$  have  $m - K_{(m)}$  and  $H - K_{(m)}$  columns, respectively. Thus, the columns of  $\mathcal{P}_2$  and  $\mathcal{Q}_2$  are singular vectors corresponding to the singular value 0, and so  $B_{(m)}^T \mathcal{P}_2 = \mathbf{0}$  and  $B_{(m)} \mathcal{Q}_2 = \mathbf{0}$ . We further partition  $\mathcal{Q}_2$  in the following way. Recall that  $\mu_1, \dots, \mu_H$  are the within-slice means defined in (2.10). By Theorem 2.1,  $\text{span}(\mu_1 - \mu, \dots, \mu_H - \mu)$  is a subspace of  $\text{span}(\Gamma_X \beta_1, \dots, \Gamma_X \beta_K)$  and therefore has dimension less than or equal to  $K \leq K_0$ . It follows from the “rank-nullity theorem” that there exists a matrix  $\mathcal{Q}_{2\circ}$  of dimension  $H \times (H - K_0)$  with orthonormal columns such that

$$(A.1) \quad (\mu_1 - \mu, \dots, \mu_H - \mu) \star (F \mathcal{Q}_{2\circ}) = \mathbf{0}.$$

Furthermore, observe that  $\mathbf{g}$  spans the null space of  $F$  and so  $(\mu_1 - \mu, \dots, \mu_H - \mu) \star (F \mathbf{g}) = 0$ . Without loss of generality, let  $\mathbf{g}$  be the last column of  $\mathcal{Q}_{2\circ}$ . Define an operator  $T: \mathcal{H} \rightarrow \mathbb{R}^m$  by

$$(A.2) \quad Tx = (\omega_1^{-1/2} \langle \psi_1, x \rangle, \dots, \omega_m^{-1/2} \langle \psi_m, x \rangle)^T, \quad x \in \mathcal{H}.$$

Applying  $T$  to both sides of (A.1), we have

$$(A.3) \quad B_{(m)} \mathcal{Q}_{2\circ} = MF \mathcal{Q}_{2\circ} = T\{(\mu_1 - \mu, \dots, \mu_H - \mu)\} F \mathcal{Q}_{2\circ} = \mathbf{0},$$

where, for convenience, the notation  $T\{(\mu_1 - \mu, \dots, \mu_H - \mu)\}$  means  $(T(\mu_1 - \mu), \dots, T(\mu_H - \mu))$ . This means  $\mathcal{Q}_{2\circ}$  is contained in the column space of  $\mathcal{Q}_2$ . Without loss of generality, we assume that  $\mathcal{Q}_2$  has the decomposition

$$(A.4) \quad \mathcal{Q}_2 = [\mathcal{Q}_{2*} | \mathcal{Q}_{2\circ}],$$

where  $\mathcal{Q}_{2*}$  is of dimension  $H \times (K_0 - K_{(m)})$ . When  $m$  is large enough so that  $K_{(m)} = K_0$ , then  $\mathcal{Q}_2 = \mathcal{Q}_{2\circ}$ .

Let

$$(A.5) \quad U_n := \mathcal{P}_2^T \widehat{B}_{(m)} \mathcal{Q}_2 = \mathcal{P}_2^T (\widehat{B}_{(m)} - B_{(m)}) \mathcal{Q}_2.$$

Also define

$$(A.6) \quad \widetilde{\boldsymbol{\vartheta}}_{h,(m)} = \frac{1}{n_h} \sum_i \boldsymbol{\eta}_{i,(m)} I(Y_i \in S_h) \quad \text{and} \quad \widetilde{M} = [\widetilde{\boldsymbol{\vartheta}}_{1,(m)}, \dots, \widetilde{\boldsymbol{\vartheta}}_{H,(m)}]_{m \times H}.$$

LEMMA 1. *Assume that  $X(t)$  has an elliptically contoured distribution satisfying (2.7). Let  $\widetilde{M}$  be defined by (A.6). We have  $\sqrt{n} \mathcal{P}_2^T \widetilde{M} G \xrightarrow{d} \mathcal{Z} \Lambda$ , where  $\mathcal{Z}$  is a  $(m - K_{(m)}) \times H$  matrix of independent  $\text{Normal}(0, 1)$  random variables,  $\Lambda = \text{diag}(\tau_1^{1/2}, \dots, \tau_H^{1/2})$  with  $\tau_h = \mathbb{E}(\Theta^2 | Y \in S_h)$ .*

PROOF. Let

$$\mathbf{u}_{n,h} = \frac{1}{n} \sum_{i=1}^n \boldsymbol{\eta}_{i,(m)} I(Y_i \in S_h), \quad p_{n,h} = \frac{1}{n} \sum_{i=1}^n I(Y_i \in S_h),$$

then  $\widetilde{\boldsymbol{\vartheta}}_{h,(m)} = \mathbf{u}_{n,h}/p_{n,h}$ , denote  $\mathbf{u}_h = \mathbb{E}(\mathbf{u}_{n,h}) = \boldsymbol{\vartheta}_{h,(m)} p_h$ . Then

$$\widetilde{M} = \left( \frac{\mathbf{u}_{n,1}}{p_{n,1}}, \dots, \frac{\mathbf{u}_{n,H}}{p_{n,H}} \right)_{m \times H}$$

and

$$M = \left( \frac{\mathbf{u}_1}{p_1}, \dots, \frac{\mathbf{u}_H}{p_H} \right)_{m \times H},$$

and so

$$\begin{aligned} & n^{1/2} (\widetilde{M} - M) \\ &= n^{1/2} \left( \frac{\mathbf{u}_{n,1}}{p_{n,1}} - \frac{\mathbf{u}_1}{p_1}, \dots, \frac{\mathbf{u}_{n,H}}{p_{n,H}} - \frac{\mathbf{u}_H}{p_H} \right) \\ &= n^{1/2} \left( \frac{\mathbf{u}_{n,1} - \mathbf{u}_1}{p_{n,1}}, \dots, \frac{\mathbf{u}_{n,H} - \mathbf{u}_H}{p_{n,H}} \right) \\ (A.7) \quad & - n^{1/2} \left( \frac{\mathbf{u}_1}{p_{n,1} p_1} (p_{n,1} - p_1), \dots, \frac{\mathbf{u}_H}{p_{n,H} p_H} (p_{n,H} - p_H) \right) \\ &= n^{1/2} \left( \frac{\mathbf{u}_{n,1} - \mathbf{u}_1}{p_1}, \dots, \frac{\mathbf{u}_{n,H} - \mathbf{u}_H}{p_H} \right) \\ & - n^{1/2} \left( \frac{\mathbf{u}_1}{p_1^2} (p_{n,1} - p_1), \dots, \frac{\mathbf{u}_H}{p_H^2} (p_{n,H} - p_H) \right) \\ & + o_p(1). \end{aligned}$$

By the central limit theorem and covariance computations, it is easy to see that the columns of  $\sqrt{n}\mathcal{P}_2^T(\widehat{M} - M)G$  are asymptotically independent, where the  $h$ th column converges in distribution to a random vector having the multivariate normal distribution

$$(A.8) \quad \text{Normal}(\mathbf{0}, \text{Var}(\mathcal{P}_2^T \boldsymbol{\eta}_{(m)} | Y \in S_h)).$$

For convenience, let  $\mathcal{V}$  denote the vector  $(\langle \beta_1, X \rangle, \dots, \langle \beta_K, X \rangle)$ . By iterative conditioning,

$$\begin{aligned} \text{Var}(\mathcal{P}_2^T \boldsymbol{\eta}_{(m)} | Y \in S_h) &= \mathbb{E}\{\text{Var}(\mathcal{P}_2^T \boldsymbol{\eta}_{(m)} | \mathcal{V}, \Theta, Y, \varepsilon) | Y \in S_h\} \\ &\quad + \text{Var}\{\mathcal{P}_2^T \mathbb{E}(\boldsymbol{\eta}_{(m)} | \mathcal{V}, \Theta, Y, \varepsilon) | Y \in S_h\} \\ &= \mathbb{E}\{\text{Var}(\mathcal{P}_2^T \boldsymbol{\eta}_{(m)} | \mathcal{V}, \Theta) | Y \in S_h\} \\ &\quad + \text{Var}\{\mathcal{P}_2^T \mathbb{E}(\boldsymbol{\eta}_{(m)} | \mathcal{V}, \Theta) | Y \in S_h\}, \end{aligned}$$

where we used the facts that  $Y$  is redundant given  $\varepsilon$  and the  $\langle \beta_k, X \rangle$ 's, and  $X$  is independent of  $\varepsilon$ . With the notation  $\check{\mathcal{V}} = (\langle \beta_1, \check{X} \rangle, \dots, \langle \beta_K, \check{X} \rangle)$ , we have

$$(A.9) \quad \begin{aligned} \text{Var}(\mathcal{P}_2^T \boldsymbol{\eta}_{(m)} | Y \in S_h) &= \mathbb{E}\{\Theta^2 \text{Var}(\mathcal{P}_2^T \check{\boldsymbol{\eta}}_{(m)} | \check{\mathcal{V}}) | Y \in S_h\} \\ &\quad + \text{Var}\{\Theta^2 \mathbb{E}(\mathcal{P}_2^T \check{\boldsymbol{\eta}}_{(m)} | \check{\mathcal{V}}) | Y \in S_h\}. \end{aligned}$$

In the following, we focus on the special case  $\check{X}$  is Gaussian. The general case is similar but requires a more careful analysis of the conditional distribution of jointly elliptically contoured random variables. Let  $\mathbf{b}$  be any column of  $\mathcal{P}_2$ . Then

$$\mathbb{E}(\mathbf{b}^T \check{\boldsymbol{\eta}}_{(m)}) = 0 \quad \text{and} \quad \mathbb{E}(\mathbf{b}^T \check{\boldsymbol{\eta}}_{(m)} \langle \beta_k, \check{X} \rangle) = \mathbf{b}^T \mathbf{b}_{k,(m)} = 0, \quad 1 \leq k \leq K.$$

Thus,  $\mathcal{P}_2^T \check{\boldsymbol{\eta}}_{(m)}$  is a vector of standard normal random variables that are independent of the  $\langle \beta_k, \check{X} \rangle$ 's. It follows from (A.9) that

$$(A.10) \quad \text{Var}(\mathcal{P}_2^T \boldsymbol{\eta}_{(m)} | Y \in S_h) = \mathbb{E}(\Theta^2 | Y \in S_h) I = \tau_h I,$$

where  $I$  is the identity matrix. The proof is complete.  $\square$

LEMMA 2. *Let  $X(t)$ ,  $\mathcal{Z}$  and  $\Lambda$  be as in Lemma 1. Then  $\sqrt{n}U_n \xrightarrow{d} Z$  where all of the entries of  $Z$  are normally distributed with mean zero and have the following properties:*

- (i) *If  $K_{(m)} = K_0$ , then  $Z \stackrel{d}{=} \mathcal{Z} \Lambda \mathcal{J}_{\mathbf{g}} \mathcal{Q}_2$ .*
- (ii) *If  $K_{(m)} < K_0$ , then  $Z$  can be partitioned as  $Z = [Z_* | Z_\circ]$  in accordance with the partition of  $\mathcal{Q}_2$  in (A.4), where  $Z_\circ \stackrel{d}{=} \mathcal{Z} \Lambda \mathcal{J}_{\mathbf{g}} \mathcal{Q}_{2\circ}$ ; furthermore, if  $X$  is Gaussian then  $Z_*$  and  $Z_\circ$  are independent, where the last column of  $Z_\circ$  is identically 0 while the rest of the entries of  $Z_\circ$  are i.i.d. standard normal.*



PROOF. First, write

$$\begin{aligned} n^{1/2}U_n &= n^{1/2}\mathcal{P}_2^T\widehat{M}\widehat{F}\mathcal{Q}_2 \\ &= n^{1/2}\mathcal{P}_2^T\{\widetilde{M}F + (\widehat{M} - \widetilde{M})F + (\widehat{M} - \widetilde{M})(\widehat{F} - F) + \widetilde{M}(\widehat{F} - F)\}\mathcal{Q}_2. \end{aligned}$$

Denote  $\bar{X}_h = \sum_i X_i I(Y_i \in S_h) / \sum_i I(Y_i \in S_h)$ . Then

$$\widehat{M} = \{\langle \widehat{\omega}_j^{-1/2} \widehat{\psi}_j, \bar{X}_h - \bar{X} \rangle\}_{j,h=1}^{m,H}, \quad \widetilde{M} = \{\langle \omega_j^{-1/2} \psi_j, \bar{X}_h - \mu \rangle\}_{j,h=1}^{m,H}.$$

Then

$$\begin{aligned} \widehat{M} - \widetilde{M} &= \{\langle \widehat{\omega}_j^{-1/2} \widehat{\psi}_j - \omega_j^{-1/2} \psi_j, \bar{X}_h - \bar{X} \rangle\}_{j,h=1}^{m,H} \\ &\quad - \{\langle \omega_j^{-1/2} \psi_j, \bar{X} - \mu \rangle\}_{j,h=1}^{m,H}. \end{aligned} \tag{A.11}$$

It follows that

$$\begin{aligned} \widehat{\psi}_j(t) - \psi_j(t) &= \sum_{\ell \neq j} \frac{\psi_\ell(t)}{\omega_j - \omega_\ell} \langle (\widehat{\Gamma}_X - \Gamma_X) \psi_\ell, \psi_j \rangle + O_p(n^{-1}), \\ \widehat{\omega}_j - \omega_j &= \langle (\widehat{\Gamma}_X - \Gamma_X) \psi_\ell, \psi_j \rangle + O_p(n^{-1}). \end{aligned} \tag{A.12}$$

These were established by (2.8) and (2.9) in Hall and Hosseini-Nasab (2006) for  $\mathcal{H} = L^2[a, b]$ . Actually, they hold for any Hilbert space  $\mathcal{H}$ ; see Eubank and Hsing (2010), Theorem 3.8.11. Since  $\widehat{\Gamma}_X - \Gamma_X = O_p(n^{-1/2})$ , these imply  $\widehat{\omega}_j = \omega_j + O_p(n^{-1/2})$  and  $\widehat{\psi}_j = \psi_j + O_p(n^{-1/2})$ . Also  $\bar{X} - \mu = O_p(n^{-1/2})$ . Thus,  $\widehat{M} - \widetilde{M} = O_p(n^{-1/2})$ . Since we also have  $\widehat{F} - F = O_p(n^{-1/2})$ , we conclude that

$$n^{1/2}\mathcal{P}_2^T(\widehat{M} - \widetilde{M})(\widehat{F} - F) = O_p(n^{-1/2}).$$

Similarly, since  $\mathcal{P}_2^T M = 0$ ,

$$n^{1/2}\mathcal{P}_2^T \widetilde{M}(\widehat{F} - F) = n^{1/2}\mathcal{P}_2^T(\widetilde{M} - M)(\widehat{F} - F) = O_p(n^{-1/2}).$$

Thus,

$$n^{1/2}U_n = n^{1/2}\mathcal{P}_2^T \widetilde{M}F\mathcal{Q}_2 + n^{1/2}\mathcal{P}_2^T(\widehat{M} - \widetilde{M})F\mathcal{Q}_2 + o_p(1). \tag{A.13}$$

To get the desired result, we break the proof into several parts. First, we establish that

$$(n^{1/2}\mathcal{P}_2^T(\widetilde{M} - M)F\mathcal{Q}_2, n^{1/2}\mathcal{P}_2^T(\widehat{M} - \widetilde{M})F\mathcal{Q}_2) \xrightarrow{d} (Z_1, Z_2), \tag{A.14}$$

where  $(Z_1, Z_2)$  are jointly normal with mean 0. By (A.11) and the fact that  $(1, \dots, 1)F = \mathbf{0}$ , we have

$$n^{1/2}(\widehat{M} - \widetilde{M})F\mathcal{Q}_2 = n^{1/2}\{\langle \widehat{\omega}_j^{-1/2} \widehat{\psi}_j - \omega_j^{-1/2} \psi_j, \bar{X}_h - \bar{X} \rangle\}_{j,h=1}^{m,H}F\mathcal{Q}_2.$$

Since  $\widehat{\omega}_i^{-1/2}\widehat{\psi}_i - \omega_i^{-1/2}\psi_i = O_p(n^{-1/2})$  and  $\bar{X}_h - \bar{X} \xrightarrow{p} \mu_h - \mu$ ,

$$(A.15) \quad n^{1/2}(\widehat{M} - \widetilde{M})F\mathcal{Q}_2 = n^{1/2}\{\langle \widehat{\omega}_j^{-1/2}\widehat{\psi}_j - \omega_j^{-1/2}\psi_j, \gamma_h \rangle\}_{j,h=1}^{m, H-K(m)} + o_p(1),$$

where

$$(\gamma_1, \dots, \gamma_{H-K(m)}) = (\mu_1 - \mu, \dots, \mu_H - \mu) \star (F\mathcal{Q}_2).$$

Note that the last  $H - K_0$  of the  $\gamma_k$ 's are equal to 0 by (A.1). In particular, if  $K(m) = K_0$  then all of the  $\gamma_h$ 's are equal to 0 and (A.14) is established with  $Z_2 = 0$  and  $Z_1 \stackrel{d}{=} \mathcal{Z}\Lambda\mathcal{T}_{\mathbf{g}}\mathcal{Q}_2$  by Lemma 1. The assertion (i) follows readily from (A.13). Below, we focus on the case  $K(m) < K_0$ . Recall that

$$\{\langle \omega_j^{-1/2}\psi_j, \gamma_h \rangle\}_{j,h=1}^{m, H-K(m)} = MF\mathcal{Q}_2 = 0,$$

which implies that

$$(A.16) \quad \{\langle \psi_j, \gamma_h \rangle\}_{j,h=1}^{m, H-K(m)} = 0.$$

By (A.15) and (A.16),

$$(A.17) \quad n^{1/2}(\widehat{M} - \widetilde{M})F\mathcal{Q}_{2*} = n^{1/2}\{\langle \omega_j^{-1/2}(\widehat{\psi}_j - \psi_j), \gamma_h \rangle\}_{j,h=1}^{m, K_0-K(m)} + o_p(1).$$

By the central limit theorem, the random element  $n^{-1/2}(\widehat{R} - R, \mathbf{u}_{n,h} - \mathbf{u}_h, p_{n,h} - p_h, h = 1, \dots, H)$  has a jointly Gaussian limit. In view of (A.7), (A.12) and (A.17), the claim in (A.14) is established by performing a linear transformation. Define the partitions  $Z_1 = [Z_{1*}|Z_{1\circ}]$  and  $Z_2 = [Z_{2*}|Z_{2\circ}]$  and so  $[Z_*|Z_{\circ}] = [Z_{1*} + Z_{2*}|Z_{1\circ}]$  since  $Z_{2\circ} = 0$ . By Lemma 1,

$$(A.18) \quad [Z_{1*}|Z_{1\circ}] \stackrel{d}{=} [\mathcal{Z}\Lambda\mathcal{J}_{\mathbf{g}}\mathcal{Q}_{2*}|\mathcal{Z}\Lambda\mathcal{T}_{\mathbf{g}}\mathcal{Q}_{2\circ}]$$

and so  $Z_{\circ} = Z_{1\circ} \stackrel{d}{=} \mathcal{Z}\Lambda\mathcal{J}_{\mathbf{g}}\mathcal{Q}_{2\circ}$ . Assume for the rest of the proof that  $X$  is Gaussian. Recall that  $\mathcal{J}_{\mathbf{g}} = I - \mathbf{g}\mathbf{g}^T$ . By the fact that  $\tau_h \equiv 1$  and the convention that  $\mathbf{g}$  is the last column of  $\mathcal{Q}_2$ , it follows from (A.18) that

$$[Z_{1*}|Z_{1\circ}] \stackrel{d}{=} [\mathcal{Z}\mathcal{Q}_{2*}|\mathcal{Z}\widetilde{\mathcal{Q}}_{2\circ}],$$

where  $\widetilde{\mathcal{Q}}_{2\circ}$  denotes the matrix whose last column contains 0's but the remaining entries are taken after  $\mathcal{Q}_{2\circ}$ . By Lemma 1,  $Z_{1*}$  and  $Z_{1\circ}$  are independent. So it remains to show that  $Z_{2*}$  and  $Z_{1\circ}$  are independent. By (A.12) and (A.16), with  $\gamma_{k\ell} := \langle \gamma_k, \psi_{\ell} \rangle$ ,

$$n^{1/2}\langle \omega_j^{-1/2}(\widehat{\psi}_j - \psi_j), \gamma_k \rangle$$

$$\begin{aligned}
&= n^{1/2} \omega_j^{-1/2} \sum_{\ell=m+1}^{\infty} \frac{\gamma_{k\ell}}{\omega_j - \omega_\ell} \int (\widehat{R} - R) \psi_\ell \psi_j + o_p(1) \\
\text{(A.19)} \quad &= n^{1/2} \omega_j^{-1/2} \sum_{\ell=m+1}^{\infty} \frac{\gamma_{k\ell}}{\omega_j - \omega_\ell} \left\{ \frac{1}{n} \sum_{i=1}^n \xi_{i\ell} \xi_{ij} \right\} + o_p(1) \\
&= n^{-1/2} \sum_{i=1}^n \left( \sum_{\ell=m+1}^{\infty} \frac{\gamma_{k\ell}}{\omega_j - \omega_\ell} \xi_{i\ell} \right) \eta_{ij} + o_p(1) \\
&=: z_{jk} + o_p(1),
\end{aligned}$$

for  $j = 1, \dots, m$ ,  $k = 1, \dots, K_0 - K_{(m)}$ . Let  $\mathbf{z}_k = (z_{1k}, \dots, z_{mk})^T$ . By (A.17) and (A.19),

$$\text{(A.20)} \quad n^{1/2} \mathcal{P}_2^T (\widehat{M} - \widetilde{M}) F \mathcal{Q}_{2*} = \mathcal{P}_2^T [\mathbf{z}_1, \dots, \mathbf{z}_{K_0 - K_{(m)}}] + o_p(1).$$

Since  $MF\mathcal{Q}_2 = 0$  and  $\mathcal{P}_2^T \mathbf{u}_h = 0$ , it follows from (A.7) that

$$\text{(A.21)} \quad n^{1/2} \mathcal{P}_2^T (\widetilde{M} - M) F \mathcal{Q}_{2\circ} = n^{1/2} \mathcal{P}_2^T \left( \frac{\mathbf{u}_{n,1}}{p_1}, \dots, \frac{\mathbf{u}_{n,H}}{p_H} \right) F \mathcal{Q}_{2\circ} + o_p(1).$$

We now compute the covariances between the components of (A.20) and (A.21). Since the components are jointly normal, our goal is to show that the covariances are all 0. Let  $\mathbf{q}$  be a column of  $\mathcal{Q}_{2\circ}$  and  $k = 1, \dots, K_0 - K_{(m)}$ . Note that

$$\mathcal{P}_2^T \mathbf{z}_k = \frac{1}{n^{1/2}} \sum_{i=1}^n \mathcal{P}_2^T \mathcal{D}_{ik} \boldsymbol{\eta}_{i,(m)},$$

where

$$\mathcal{D}_{ik} = \text{diag} \left( \sum_{\ell=m+1}^{\infty} \frac{\gamma_{\ell k}}{\omega_j - \omega_\ell} \xi_{i\ell}, j = 1, \dots, m \right).$$

Since  $\mathbb{E}(\mathcal{P}_2^T \mathbf{z}_k) = 0$ ,

$$\begin{aligned}
&\text{Cov} \left( n^{1/2} \mathcal{P}_2^T \left( \frac{\mathbf{u}_{n,1}}{p_1}, \dots, \frac{\mathbf{u}_{n,H}}{p_H} \right) F \mathbf{q}, \mathcal{P}_2^T \mathbf{z}_k \right) \\
&= \mathbb{E} \left( n^{1/2} \mathcal{P}_2^T \left( \frac{\mathbf{u}_{n,1}}{p_1}, \dots, \frac{\mathbf{u}_{n,H}}{p_H} \right) F \mathbf{q} \mathbf{z}_k^T \mathcal{P}_2 \right) \\
&= \mathbb{E} \left( n^{1/2} \left( \frac{\mathcal{P}_2^T \mathbf{u}_{n,1} \mathbf{z}_k^T \mathcal{P}_2}{p_1}, \dots, \frac{\mathcal{P}_2^T \mathbf{u}_{n,H} \mathbf{z}_k^T \mathcal{P}_2}{p_H} \right) \star (F \mathbf{q}) \right).
\end{aligned}$$

Let  $\mathcal{V}$  be as defined in the proof of Lemma 1. By the same conditioning argument employed there,

$$\text{(A.22)} \quad \mathbb{E}(n^{1/2} \mathcal{P}_2^T \mathbf{u}_{n,h} \mathbf{z}_k^T \mathcal{P}_2) = \mathbb{E}[\mathbb{E}\{\mathcal{P}_2^T \boldsymbol{\eta}_{(m)} \boldsymbol{\eta}_{(m)}^T \mathcal{D}_k \mathcal{P}_2 | \mathcal{V}\} I(Y \in S_h)],$$

where the sub-index  $i$  is suppressed from the symbols on the right-hand side since it suffices to deal with a generic process  $(X, Y)$  in computing expectations. Note that  $\mathcal{P}_2^T \boldsymbol{\eta}_{(m)}$  and  $\mathcal{V}$  are independent,  $\boldsymbol{\eta}_{(m)}$  and  $\mathcal{D}_k$  are independent, and  $\boldsymbol{\eta}_{(m)}, \mathcal{V}, \mathcal{D}_k$  are normally distributed with mean zero. Then it is easy to conclude from (A.22) that

$$\begin{aligned}
 & \mathbb{E}(n^{1/2} \mathcal{P}_2^T \mathbf{u}_{n,h} \mathbf{z}_k^T \mathcal{P}_2) \\
 (A.23) \quad &= \mathbb{E}(\mathcal{P}_2^T \boldsymbol{\eta}_{(m)} \boldsymbol{\eta}_{(m)}^T \mathcal{P}_2) \mathbb{E}[\mathbb{E}\{\mathcal{P}_2 \mathcal{D}_k \mathcal{P}_2 | \mathcal{V}\} I(Y \in S_h)] \\
 &= \mathbb{E}[\mathbb{E}\{\mathcal{P}_2 \mathcal{D}_k \mathcal{P}_2 | \mathcal{V}\} I(Y \in S_h)].
 \end{aligned}$$

By the property of the normal distribution, each (diagonal) element of  $\mathbb{E}\{\mathcal{D}_k | \mathcal{V}\}$  can be written as  $\sum_{j=1}^K c_j \langle \beta_j, X - \mu \rangle$  for some  $c_j, 1 \leq j \leq K$ . For convenience, denote  $\mathbb{E}\{\mathcal{D}_k | \mathcal{V}\}$  as  $T(X - \mu)$  where  $T$  is a linear functional. Thus,

$$(A.24) \quad \mathbb{E}[\mathbb{E}\{\mathcal{P}_2^T \mathcal{D}_k \mathcal{P}_2 | \mathcal{V}\} I(Y \in S_h)] = p_h \mathcal{P}_2^T T(\mu_h - \mu) \mathcal{P}_2.$$

As a result,

$$\begin{aligned}
 & \mathbb{E}\left(n^{1/2} \left( \frac{\mathcal{P}_2^T \mathbf{u}_{n,1} \mathbf{z}_k^T \mathcal{P}_2}{p_1}, \dots, \frac{\mathcal{P}_2^T \mathbf{u}_{n,H} \mathbf{z}_k^T \mathcal{P}_2}{p_H} \right) \star (F\mathbf{q})\right) \\
 &= (\mathcal{P}_2^T T(\mu_1 - \mu) \mathcal{P}_2, \dots, \mathcal{P}_2^T T(\mu_H - \mu) \mathcal{P}_2) \star (F\mathbf{q}) \\
 &= \mathcal{P}_2^T ((T(\mu_1 - \mu), \dots, T(\mu_H - \mu)) \star (F\mathbf{q})) \mathcal{P}_2 = 0,
 \end{aligned}$$

by (A.1). This shows that the covariances between the components of (A.20) and (A.21) are all equal to 0, and concludes the proof that  $Z_{1\circ}$  and  $Z_{2*}$  are independent.  $\square$

**PROOF OF PROPOSITION 3.2.** Assume for convenience that  $Z_1$  has full column rank. If this is not the case, a slight modification of the proof below suffices. Denote the  $j$ th column of  $Z$  as  $\mathbf{z}_j$ , and construct orthonormal vectors by applying the Gram-Schmidt orthonormalization to the columns of  $Z$ :

$$\mathbf{v}_1 = \frac{\mathbf{z}_1}{\|\mathbf{z}_1\|}, \quad \mathbf{v}_j = \frac{(I - \Pi_{j-1})\mathbf{z}_j}{\|(I - \Pi_{j-1})\mathbf{z}_j\|}, \quad j = 2, \dots, \min(p, q),$$

where  $\Pi_{j-1} = [\mathbf{v}_1, \dots, \mathbf{v}_{j-1}][\mathbf{v}_1, \dots, \mathbf{v}_{j-1}]^T$  is the projection matrix to the space spanned by  $\mathbf{z}_1, \dots, \mathbf{z}_{j-1}$ . The following properties can be verified:

- (a)  $\mathbf{v}_j^T \mathbf{z}_k = 0$  for all pairs  $k < j$ . This is the result of the construction of the  $\mathbf{v}_j$ 's.
- (b)  $\mathbf{v}_j$  is independent of  $\mathbf{z}_k$  for  $k > \max(j, r)$ . This follows from the assumption on  $Z_2$ .

- (c)  $\mathbf{v}_j^T \mathbf{z}_k \sim \text{Normal}(0, 1)$  for  $k > \max(j, r)$ . The proof of this is easy: by (b) and the fact that  $\|\mathbf{v}_j\| = 1$ ,  $(\mathbf{v}_j^T \mathbf{z}_k | \mathbf{v}_j) \sim \text{Normal}(0, 1)$ ; since this conditional distribution does not depend on  $\mathbf{v}_j$ , it is also the marginal distribution.
- (d)  $\mathbf{v}_j^T \mathbf{z}_k$  and  $\mathbf{v}_{j'}^T \mathbf{z}_{k'}$  are independent if  $k > \max(j, r)$  and  $k' > \max(j', r)$ . The proof is as follows. First for the case  $j, j', k < k'$ , we have

$$\begin{aligned} \mathbb{P}(\mathbf{v}_j^T \mathbf{z}_k \leq x, \mathbf{v}_{j'}^T \mathbf{z}_{k'} \leq y) &= \mathbb{E}[\mathbb{P}(\mathbf{v}_j^T \mathbf{z}_k \leq x, \mathbf{v}_{j'}^T \mathbf{z}_{k'} \leq y | \mathbf{v}_j, \mathbf{v}_{j'}, \mathbf{z}_k)] \\ &= \mathbb{E}[I(\mathbf{v}_j^T \mathbf{z}_k \leq x) \mathbb{P}(\mathbf{v}_{j'}^T \mathbf{z}_{k'} \leq y | \mathbf{v}_{j'})] \\ &= \Phi(x) \Phi(y), \end{aligned}$$

where the last step follows from (c). Next for  $j, j' < k = k'$  and  $j \neq j'$ , we have

$$\begin{aligned} \mathbb{P}(\mathbf{v}_j^T \mathbf{z}_k \leq x, \mathbf{v}_{j'}^T \mathbf{z}_k \leq y) &= \mathbb{E}[\mathbb{P}(\mathbf{v}_j^T \mathbf{z}_k \leq x, \mathbf{v}_{j'}^T \mathbf{z}_k \leq y | \mathbf{v}_j, \mathbf{v}_{j'})] = \Phi(x) \Phi(y) \\ \text{since } \mathbf{v}_j^T \mathbf{v}_{j'} &= 0. \end{aligned}$$

- (e)  $\mathbf{v}_j^T \mathbf{z}_j = \|(I - \Pi_{j-1})\mathbf{z}_j\|$  for  $j \geq r+1$  is the square root of a  $\chi_{p-j+1}^2$  variable, and it is independent of any  $\mathbf{v}_{j'}^T \mathbf{z}_{k'}$  with  $k' > j' \geq r$ . The claims can be easily verified using conditioning arguments similar to those in (c) and (d).
- (f)  $\mathbf{v}_j^T \mathbf{z}_j$  is independent of  $\mathbf{v}_{j'}^T \mathbf{z}_{j'}$ , for  $j, j' \geq r+1$  and  $j \neq j'$ . This can be verified by checking the independence between  $(I - \Pi_{j-1})\mathbf{z}_j$  and  $(I - \Pi_{j'-1})\mathbf{z}_{j'}$ .

Based on (a)–(f), we conclude that the entries in  $[\mathbf{v}_1, \dots, \mathbf{v}_{\min(p,q)}]^T Z$  have the following properties: all entries below the diagonal are zero; all entries in the last  $q - r$  columns and on and above the diagonal are independent, where those above the diagonal are distributed as standard normal and the square of the  $j$ th diagonal element is distributed as  $\chi_{p-j+1}^2$ .

Notice that if  $p \leq q$ , the  $\mathbf{v}_j$ 's defined above already constitute a basis for  $\mathbb{R}^p$ . If  $p > q$ , we can define  $\mathbf{v}_j$ ,  $j = q+1, \dots, p$ , such that they are orthogonal to all columns of  $Z$ , and to each other. Define  $V_r = [\mathbf{v}_{r+1}, \dots, \mathbf{v}_p]$ . By the nature of eigenvalues,

$$\begin{aligned} &\sum_{j=r+1}^p \lambda_j(ZZ^T) \\ &= \min_{\Phi} \{\text{tr}(\Phi^T ZZ^T \Phi), \\ &\quad \Phi \text{ is a } p \times (p-r) \text{ matrix with orthonormal columns}\} \\ &\leq \text{tr}(V_r^T ZZ^T V_r) = \sum_{j=r+1}^p \mathbf{v}_j^T Z Z^T \mathbf{v}_j. \end{aligned} \tag{A.25}$$

It follows from the summary above that the last expression is a sum of independent  $\chi^2$  random variable, and a simple calculation shows that the total degrees of freedom is  $(p-r)(q-r)$ .  $\square$

**PROOF OF THEOREM 3.1.** To study the smallest  $m - K_0$  eigenvalues of  $\hat{V}_{(m)}$ , we can equivalently study the smallest squared  $m - K_0$  singular values of  $\hat{B}_{(m)}$ . By the asymptotic theory described in Dauxois, Pousse and Romain (1982) and Hall and Hosseini-Nasab (2006), it is straightforward to show that  $\sqrt{n}(\hat{B}_{(m)} - B_{(m)})$  converges in distribution. By Theorem 4.1 in Eaton and Tyler (1994), the pairwise difference between the smallest  $m - K_{(m)}$  singular values of  $\hat{B}_{(m)}$  and the singular values of  $U_n = \mathcal{P}_2^T \hat{B}_{(m)} \mathcal{Q}_2$  is  $O_p(n^{-3/4})$ . So, for  $K_{(m)} = K_0$ , the smallest  $m - K_0$  eigenvalues of  $\hat{V}_{(m)}$  are approximated by the complete set of eigenvalues of  $U_n U_n^T$ , while, for  $K_{(m)} < K_0$ , the smallest  $m - K_0$  eigenvalues of  $\hat{V}_{(m)}$  are only approximated by a subset of eigenvalues of  $U_n U_n^T$ . We consider the two cases in more details below.

(i) For  $K_{(m)} = K_0$ , we will prove (3.4) from which (3.2) follows easily. It follows that  $\mathcal{Q}_2 = \mathcal{Q}_{2\circ}$  and

$$\mathcal{T}_{K_0, (m)} = n \operatorname{tr}(U_n U_n^T) + o_p(1).$$

By (i) of Lemma 2,

$$(A.26) \quad \mathcal{T}_{K_0, (m)} \xrightarrow{d} \operatorname{tr}(\mathcal{Z} \Lambda \Xi \Lambda \mathcal{Z}^T),$$

where  $\mathcal{Z}$  is as given in Lemma 1 and  $\Xi := \mathcal{J}_{\mathbf{g}} \mathcal{Q}_2 \mathcal{Q}_2^T \mathcal{J}_{\mathbf{g}}$ . It is easy to see that  $\mathcal{J}_{\mathbf{g}}$  and  $\mathcal{Q}_2 \mathcal{Q}_2^T$  are projection matrices with rank  $H - 1$  and  $H - K_0$ , respectively. Since  $\mathbf{g}$  is a column of  $\mathcal{Q}_2$ , we have  $\mathcal{Q}_2 \mathcal{Q}_2^T \mathbf{g} = \mathbf{g}$ . As a result,

$$\Xi = \mathcal{J}_{\mathbf{g}} \mathcal{Q}_2 \mathcal{Q}_2^T \mathcal{J}_{\mathbf{g}} = \mathcal{Q}_2 \mathcal{Q}_2^T - \mathbf{g} \mathbf{g}^T,$$

which is a projection matrix with rank and trace equal to  $H - 1 - K_0$ . Since  $\Lambda$  is full rank,  $\operatorname{Rank}(\Lambda \Xi \Lambda) = \operatorname{Rank}(\Xi) = H - K_0 - 1$ . Let  $A \Delta A^T$  be the eigen decomposition of  $\Lambda \Xi \Lambda$  where the column of  $A$  are the orthonormal eigenvectors of  $\Lambda \Xi \Lambda$  and  $\Delta = \operatorname{diag}\{\delta_1, \dots, \delta_{H-K_0-1}\}$  contains the positive eigenvalues. Write

$$\operatorname{tr}(Z \Lambda \Xi \Lambda Z^T) = \sum_{i=1}^{m-K_0} \mathbf{z}_i \Lambda \Xi \Lambda \mathbf{z}_i^T = \sum_{i=1}^{m-K_0} \mathbf{z}_i A \Delta A^T \mathbf{z}_i^T = \sum_{i=1}^{m-K_0} \sum_{k=1}^{H-K_0-1} \delta_k \chi_{i,k}^2,$$

where  $\mathbf{z}_i$  is the  $i$ th row vector of  $Z$ , and  $\chi_{i,k}^2$  is the  $k$ th element of  $\mathbf{z}_i A$ . Clearly, the  $\chi_{i,k}^2$  are i.i.d.  $\chi^2$  random variables with degree 1.

(ii) For  $K_{(m)} < K_0$ , it follows that

$$\begin{aligned}
\mathcal{T}_{K_0, (m)} &= n \times \sum_{j=K_0+1}^m \lambda_j(\widehat{B}_{(m)} \widehat{B}_{(m)}^T) \\
&= n \times \sum_{j=K_0-K_{(m)}+1}^{m-K_{(m)}} \lambda_j(U_n U_n^T) + o_p(1) \\
&\xrightarrow{d} \sum_{j=K_0-K_{(m)}+1}^{m-K_{(m)}} \lambda_j(Z Z^T)
\end{aligned}$$

by Lemma 2. Since the last column of  $Z$  is identically zero, the last expression is equal to

$$\sum_{j=K_0-K_{(m)}+1}^{m-K_{(m)}} \lambda_j\{(Z_*, Z_{\circ}^{[-1]})(Z_*, Z_{\circ}^{[-1]})^T\},$$

where  $Z_{\circ}^{[-1]}$  denotes the matrix  $Z_{\circ}$  minus the last column. We apply Proposition 3.2, with  $Z_1 = Z_*$ ,  $Z_2 = Z_{\circ}^{[-1]}$ ,  $p = m - K_{(m)}$ ,  $q = H - K_{(m)} - 1$  and  $r = K_0 - K_{(m)}$ , to obtain the desired result.  $\square$

### A.2. Proof of Theorem 3.3.

LEMMA 3.  $W_{(m)}$  has the same column space as  $B_{(m)}$ .

PROOF. By definition,  $W_{(m)} = B_{(m)} \Lambda (\Lambda \mathcal{J}_{\mathbf{g}} \Lambda)^{-}$ , therefore the column space of  $W_{(m)}$  is contained in that of  $B_{(m)}$ . Suppose the column rank of  $W_{(m)}$  is strictly less than that of  $B_{(m)}$ . Then there exists a nonzero vector  $\mathbf{x} \in \mathbb{R}^m$  such that  $\mathbf{x}^T W_{(m)} = \mathbf{0}$  but  $\mathbf{x}^T B_{(m)} \neq \mathbf{0}$ . Since  $B_{(m)} \mathbf{g} = \mathbf{0}$ ,  $B_{(m)} \mathcal{J}_{\mathbf{g}} = B_{(m)}$  and so

$$(A.27) \quad \mathbf{0} = \mathbf{x}^T W_{(m)} = \mathbf{x}^T B_{(m)} \Lambda^{-1} (\Lambda \mathcal{J}_{\mathbf{g}} \Lambda) (\Lambda \mathcal{J}_{\mathbf{g}} \Lambda)^{-}.$$

Observe that  $\Lambda^{-1} \mathbf{g}$  spans the null space of  $\Lambda \mathcal{J}_{\mathbf{g}} \Lambda$ . Since  $\mathbf{x}^T B_{(m)} \neq \mathbf{0}$ , we conclude that  $\mathbf{x}^T B_{(m)} \Lambda^{-1} = \delta (\Lambda^{-1} \mathbf{g})^T$  for some constant  $\delta$ . Thus,  $\mathbf{x}^T B_{(m)} = \delta \mathbf{g}^T$ . Since  $B_{(m)} \mathbf{g} = \mathbf{0}$ , it follows from (A.27) that  $\|\mathbf{x}^T B_{(m)}\|^2 = \mathbf{x}^T B_{(m)} B_{(m)}^T \mathbf{x} = \delta \mathbf{g}^T B_{(m)}^T \mathbf{x} = \mathbf{0}$ , which leads to a contradiction to the assumption that  $\mathbf{x}^T B_{(m)} \neq \mathbf{0}$ . The only possibility left is that the column space of  $W_{(m)}$  is the same as  $B_{(m)}$ .  $\square$

PROOF OF THEOREM 3.3. As mentioned in the proof of Theorem 3.1,  $\widehat{B}_{(m)} = B_{(m)} + O_p(n^{-1/2})$ , which leads to  $\widehat{V}_{(m)} = V_{(m)} + O_p(n^{-1/2})$  and  $\widehat{\mathcal{P}}_2 \widehat{\mathcal{P}}_2^T =$

$\mathcal{P}_2 \mathcal{P}_2^T + O_p(n^{-1/2})$ . By (3.5) and (A.10), we have  $\widehat{\tau}_h = \tau_h + O_p(n^{-1/2})$ . Therefore,  $\widehat{W}_{(m)}$  is a root  $n$  consistent estimator of  $W_{(m)}$ . The rest of the proof will follow the same general structure as that of (i) of Theorem 3.1. Suppose  $W_{(m)}$  has the singular-value decomposition

$$W_{(m)} = \mathcal{R} \begin{pmatrix} \widetilde{D} & 0 \\ 0 & 0 \end{pmatrix} \mathcal{S}^T,$$

where  $\mathcal{R}$  and  $\mathcal{S}$  are, respectively,  $m \times m$  and  $H \times H$  orthonormal matrices, and  $\widetilde{D} = \text{diag}(\lambda_1^{1/2}(\Sigma_{(m)}), \dots, \lambda_{K_{(m)}}^{1/2}(\Sigma_{(m)}))$ . As before, consider the partition  $\mathcal{R} = [\mathcal{R}_1 | \mathcal{R}_2]$  and  $\mathcal{S} = [\mathcal{S}_1 | \mathcal{S}_2]$  where  $\mathcal{R}_1$  and  $\mathcal{S}_1$  have  $K_{(m)}$  columns, and  $\mathcal{R}_2$  and  $\mathcal{S}_2$  have  $m - K_{(m)}$  and  $H - K_{(m)}$  columns, respectively. By Lemma 3,  $B_{(m)}$  and  $W_{(m)}$  have the same column space, therefore we can take  $\mathcal{R}_2 = \mathcal{P}_2$  without loss of generality. Similar to the definition of  $\mathcal{Q}_2$ , we proceed to construct  $\mathcal{S}_2$ . Again, since  $\text{span}(\mu_1 - \mu, \dots, \mu_H - \mu)$  has dimension less than or equal to  $K_0$ , there exists a matrix  $\mathcal{S}_{2\circ}$  with dimension  $H \times (H - K_0)$  and orthonormal columns such that

$$(\mu_1 - \mu, \dots, \mu_H - \mu) \star (F\Lambda(\Lambda \mathcal{J}_{\mathbf{g}}\Lambda)^{-} \mathcal{S}_{2\circ}) = \mathbf{0}.$$

Observe that  $(\mu_1 - \mu, \dots, \mu_H - \mu) \star F\Lambda(\Lambda \mathcal{J}_{\mathbf{g}}\Lambda)^{-} \Lambda^{-1} \mathbf{g} = 0$  since  $\Lambda^{-1} \mathbf{g}$  spans the null space of  $\Lambda \mathcal{J}_{\mathbf{g}}\Lambda$ . Without loss of generality, let  $\Lambda^{-1} \mathbf{g}$  be the last column of  $\mathcal{S}_{2\circ}$ . Let  $T$  be as defined in (A.2). As in (A.3), we obtain

$$\begin{aligned} W_{(m)} \mathcal{S}_{2\circ} &= MF\Lambda(\Lambda \mathcal{J}_{\mathbf{g}}\Lambda)^{-} \mathcal{S}_{2\circ} \\ &= T\{(\mu_1 - \mu, \dots, \mu_H - \mu)\} F\Lambda(\Lambda \mathcal{J}_{\mathbf{g}}\Lambda)^{-} \mathcal{S}_{2\circ} \\ &= \mathbf{0}. \end{aligned}$$

Since  $K_{(m)} = K_0$ , we can, and will, take  $\mathcal{S}_2$  to be  $\mathcal{S}_{2\circ}$ . Again, by Theorem 4.1 in Eaton and Tyler (1994), the smallest  $m - K_{(m)}$  singular values of  $\widehat{W}_{(m)}$  are asymptotically equivalent to those of  $U_n^* := \mathcal{P}_2 \widehat{W}_{(m)} \mathcal{S}_2$ , so that we have

$$\mathcal{T}_{K_0, (m)}^* = n \text{tr}\{U_n^* (U_n^*)^T\} + o_p(1).$$

Let  $\mathcal{F} = G \mathcal{J}_{\mathbf{g}}\Lambda(\Lambda \mathcal{J}_{\mathbf{g}}\Lambda)^{-}$ ,  $\widehat{\mathcal{F}} = \widehat{G} \mathcal{J}_{\widehat{\mathbf{g}}} \widehat{\Lambda}(\widehat{\Lambda} \mathcal{J}_{\widehat{\mathbf{g}}} \widehat{\Lambda})^{-}$ . Similar to Lemma 2,

$$n^{1/2} U_n^* = n^{1/2} \mathcal{P}_2^T \widehat{M} \widehat{\mathcal{F}} \mathcal{S}_2 = n^{1/2} \{ \mathcal{P}_2^T \widetilde{M} \widetilde{\mathcal{F}} \mathcal{S}_2 + \mathcal{P}_2^T (\widehat{M} - \widetilde{M}) \widehat{\mathcal{F}} \mathcal{S}_2 \} + o_p(1).$$

By arguments similar to those in the proof of Lemma 2, we have  $n^{1/2} \mathcal{P}_2^T (\widehat{M} - \widetilde{M}) \widehat{\mathcal{F}} \mathcal{S}_2 = o_p(1)$ . Let  $\Xi^* = \Lambda \mathcal{J}_{\mathbf{g}}\Lambda(\Lambda \mathcal{J}_{\mathbf{g}}\Lambda)^{-} \mathcal{S}_2 \mathcal{S}_2^T (\Lambda \mathcal{J}_{\mathbf{g}}\Lambda)^{-} \Lambda \mathcal{J}_{\mathbf{g}}\Lambda$ . Thus,

$$\mathcal{T}_{K_0, (m)}^* = n \text{tr}(\mathcal{P}_2^T \widetilde{M} \widetilde{\mathcal{F}} \mathcal{S}_2 \mathcal{S}_2^T \widetilde{\mathcal{F}}^T \widetilde{M}^T \mathcal{P}_2) + o_p(1) \xrightarrow{d} \text{tr}(Z \Xi^* Z^T)$$



by Lemma 1, where  $Z$  is  $(m - K_0) \times H$  matrix with independent  $\text{Normal}(0, 1)$  entries. By the fact that  $\Lambda \mathcal{J}_{\mathbf{g}} \Lambda (\Lambda \mathcal{J}_{\mathbf{g}} \Lambda)^{-}$  is a projection matrix with only one null vector  $\Lambda^{-1} \mathbf{g}$  and the assumption that  $\Lambda^{-1} \mathbf{g}$  is a column of  $S_2$ , it easily follows that  $\Xi^*$  is a projection matrix with trace equal to  $H - 1 - K_0$ . Therefore,  $\text{tr}(Z \Xi^* Z^T)$  is distributed as  $\chi_{(m-K_0) \times (H-K_0-1)}^2$ .  $\square$

**A.3. Proof of Theorem 3.4.** Some of the variables and matrices introduced in earlier sections depend on  $m$ , and we will add the subscript  $(m)$  to those quantities to emphasize this dependence in the proof. Consider the singular-value decomposition

$$B_{(m)} = \mathcal{P}_{(m)} \begin{pmatrix} D_{(m)} & 0 \\ 0 & 0 \end{pmatrix} \mathcal{Q}_{(m)}^T,$$

where  $\mathcal{P}_{(m)}$  and  $\mathcal{Q}_{(m)}$  have the same partition as before (see the proof of Theorem 3.1):  $\mathcal{P}_{(m)} = [\mathcal{P}_{1,(m)} | \mathcal{P}_{2,(m)}]$  and  $\mathcal{Q}_{(m)} = [\mathcal{Q}_{1,(m)} | \mathcal{Q}_{2,(m)}]$ . The nonuniqueness of  $\mathcal{P}_{2,(m)}$  and  $\mathcal{Q}_{2,(m)}$  allows us to construct  $\mathcal{P}_{2,(m)}$  and  $\mathcal{Q}_{2,(m)}$  in a particular way, as follows. It will be easier to think about the case where  $K_{(m)} = K_0$  for  $m$  large enough. We will henceforth make this assumption even though it is not necessary for the result to hold. Thus, there exists an ascending sequence  $0 < m_1 < m_2 < \dots < m_{K_0} < \infty$ , such that

$$m_j = \min\{m, K_{(m)} \geq j\}, \quad j = 1, \dots, K_0,$$

which are the instances where the rank of  $B_{(m)}$  changes.

We first construct  $\mathcal{Q}_{2,(m)}$  whose columns span the null row space of  $B_{(m)}$ . Let  $\mathcal{Q}_{2\circ}$  be as in the proof of Theorem 3.1. Define a sequence of orthonormal vectors  $\mathbf{q}_j, j = 1, \dots, K_0$  by backward induction:

$$\begin{aligned} B_{(m_{K_0-1})} \mathbf{q}_{K_0} &= \mathbf{0} \quad \text{and} \quad \mathcal{Q}_{2\circ}^T \mathbf{q}_{K_0} = \mathbf{0}; \\ B_{(m_{j-1})} \mathbf{q}_j &= \mathbf{0} \end{aligned}$$

and

$$\begin{aligned} [\mathbf{q}_{j+1}, \dots, \mathbf{q}_{K_0}, \mathcal{Q}_{2\circ}]^T \mathbf{q}_j &= \mathbf{0}, \quad j = K_0 - 1, \dots, 2; \\ [\mathbf{q}_2, \dots, \mathbf{q}_{K_0}, \mathcal{Q}_{2\circ}]^T \mathbf{q}_1 &= \mathbf{0}. \end{aligned}$$

Such a sequence of  $\mathbf{q}_j$ 's clearly exist. Define

$$(A.28) \quad \mathcal{Q}_{2,(m)} = \begin{cases} [\mathbf{q}_{K_{(m)}+1}, \dots, \mathbf{q}_{K_0}, \mathcal{Q}_{2\circ}], & m < m_{K_0}, \\ \mathcal{Q}_{2\circ}, & m \geq m_{K_0}. \end{cases}$$

Thus,  $\mathcal{Q}_{2,(m+1)} = \mathcal{Q}_{2,(m)}$  if  $K_{(m+1)} = K_{(m)}$ , otherwise  $\mathcal{Q}_{2,(m+1)}$  equals  $\mathcal{Q}_{2,(m)}$  minus the first column.

We next construct  $\mathcal{P}_{2,(m)}$ , a matrix of dimension  $m \times (m - K_{(m)})$ , whose columns generate the null column space of  $B_{(m)}$ . To do that, we start with

$m = K_0 + 1$  for which we will just make an arbitrary choice of  $\mathcal{P}_{2,(m)}$  that works. Suppose we have defined  $\mathcal{P}_{2,(m)}$  for some  $m$ . If  $K_{(m+1)} = K_{(m)} + 1$ , let

$$\mathcal{P}_{2,(m+1)} = \begin{bmatrix} \mathcal{P}_{2,(m)} \\ \mathbf{0}_{m-K_{(m)}}^T \end{bmatrix};$$

if  $K_{(m+1)} = K_{(m)}$ , let

$$\mathcal{P}_{2,(m+1)} = (\mathcal{P}_{21,(m+1)}, \mathbf{v}_{m+1}) \quad \text{where } \mathcal{P}_{21,(m+1)} = \begin{pmatrix} \mathcal{P}_{2,(m)} \\ \mathbf{0}_{m-K_{(m)}}^T \end{pmatrix}, \quad (\text{A.29})$$

where  $\mathbf{v}_{m+1}$  is a new null singular column vector in  $\mathbb{R}^{m+1}$ . Thus, a whole sequence of  $\mathcal{P}_{2,(m)}$  can be defined recursively in this manner.

We now briefly summarize some of the key points in the proof of Theorem 3.1. For each  $m \geq K_0 + 1$ , there exists a Gaussian random matrix  $Z_{(m)}$  such that

$$\begin{aligned} (\text{A.30}) \quad & \mathcal{P}_{2,(m)}^T \hat{B}_{(m)} \mathcal{Q}_{2,(m)} \xrightarrow{d} Z_{(m)} \quad \text{and} \\ & \mathcal{T}_{K_0,(m)} \xrightarrow{d} \sum_{j=K_0-K_{(m)}+1}^{m-K_{(m)}} \lambda_j(Z_{(m)} Z_{(m)}^T); \end{aligned}$$

if we write  $Z_{(m)} = [Z_{*,(m)} | Z_{\circ,(m)}]$  where  $Z_{*,(m)}$  contains the first  $K_0 - K_{(m)}$  columns of  $Z_{(m)}$ , then  $Z_{*,(m)}$  is independent of  $Z_{\circ,(m)}$ , and  $Z_{\circ,(m)}$  contains independent  $\text{Normal}(0, 1)$  random variables except the last column which contains zeros. The proof of Proposition 3.2 shows that there exist orthonormal vectors  $\phi_{1,(m)}, \dots, \phi_{m-K_0,(m)}$  in  $\mathbb{R}^{m-K_{(m)}}$  that are orthogonal to the columns of  $Z_{*,(m)}$  such that

$$(\text{A.31}) \quad \mathcal{X}_{(m)} := \sum_{j=1}^{m-K_0} \phi_{j,(m)}^T Z_{(m)} Z_{(m)}^T \phi_{j,(m)} \sim \chi_{(m-K_0) \times (H-K_0-1)}^2.$$

Note that the  $\phi_{j,(m)}$ 's are obtained by relabeling the  $\mathbf{v}_j$ 's in that proof. By (A.30) and (A.31), using the notion of (A.25), we conclude that  $\mathcal{T}_{K_0,(m)}$  is asymptotically bounded by  $\mathcal{X}_{(m)}$ . Thus, we have the desired stochastic bound for the first term,  $m = K_0 + 1$ , but so far there is nothing new. To define  $\mathcal{X}_{(m+1)}$ , we proceed in a similar manner by identifying a set of orthonormal vectors  $\phi_{j,(m+1)}, j = 1, \dots, (m+1) - K_0$ . As we will see, the specific choice of  $\mathcal{P}_{2,(m)}$  and  $\mathcal{Q}_{2,(m)}$  that was made enables us to directly relate  $Z_{(m)}$  and  $Z_{(m+1)}$  in a probability space, and, consequently, the two bounds as well.

Consider the two situations  $K_{(m+1)} = K_{(m)} + 1$  and  $K_{(m+1)} = K_{(m)}$  separately.

*Case 1:*  $K_{(m+1)} = K_{(m)} + 1$ . In view of the relationship between  $(\mathcal{P}_{2,(m)}, \mathcal{Q}_{2,(m)})$  and  $(\mathcal{P}_{2,(m+1)}, \mathcal{Q}_{2,(m+1)})$ , it is easy to see that  $Z_{(m+1)}$  is equal to  $Z_{(m)}$  less the first column. Below we denote the  $k$ th column of  $Z_{*,(m)}$  by  $\mathbf{z}_{k,(m)}$ . Define

$$\phi_{j,(m+1)} = \begin{cases} \phi_{j,(m)}, & j = 1, \dots, m - K_0, \\ \frac{(I - \Pi)\mathbf{z}_{1,(m)}}{\|(I - \Pi)\mathbf{z}_{1,(m)}\|}, & j = (m + 1) - K_0, \end{cases}$$

where  $\Pi$  is the projection matrix onto  $\text{span}\{\mathbf{z}_{2,(m)}, \dots, \mathbf{z}_{K_0-K_{(m)},(m)}\}$ . Observe that the vectors  $\phi_{j,(m+1)}, j = 1, \dots, (m + 1) - K_0$ , are orthonormal. Define

$$(A.32) \quad \mathcal{X}_{(m+1)} = \sum_{j=1}^{(m+1)-K_0} \phi_{j,(m+1)}^T Z_{(m+1)} Z_{(m+1)}^T \phi_{j,(m+1)} = \mathcal{X}_{(m)} + \mathcal{X},$$

where  $\mathcal{X} = \phi_{(m+1)-K_0,(m+1)}^T Z_{(m+1)} Z_{(m+1)}^T \phi_{(m+1)-K_0,(m+1)}$ . Note that

$$\phi_{(m+1)-K_0,(m+1)} \in \text{span}^\perp\{\mathbf{z}_{2,(m)}, \dots, \mathbf{z}_{K_0-K_{(m)},(m)}\}$$

and is independent of  $Z_{\text{o},(m)}$ . The conditioning arguments in the proof of Proposition 3.2 can be applied to conclude that  $\mathcal{X} \sim \chi_{H-K_0-1}^2$  and is independent of  $\mathcal{X}_{(m)}$ . As a result,  $\mathcal{T}_{K_0,(m)}$  and  $\mathcal{T}_{K_0,(m+1)}$  are jointly asymptotically bounded by  $\mathcal{X}_{(m)}$  and  $\mathcal{X}_{(m)} + \mathcal{X}$ .

*Case 2:*  $K_{(m+1)} = K_{(m)}$ . In this case,

$$Z_{(m+1)} = \begin{bmatrix} Z_{(m)} \\ \mathbf{w}^T \end{bmatrix},$$

where  $\mathbf{w}^T$  is the limit of  $\mathbf{v}_{m+1}^T \hat{B}_{(m+1)} \mathcal{Q}_{2,(m)}$ ; see (A.29). Define

$$\phi_{j,(m+1)} = \begin{cases} \begin{bmatrix} \phi_{j,(m)} \\ 0 \end{bmatrix}, & j = 1, \dots, m - K_0, \\ \frac{(I - \Pi)\mathbf{e}}{\|(I - \Pi)\mathbf{e}\|}, & j = (m + 1) - K_0, \end{cases}$$

where, in this case,  $\Pi$  is the projection matrix onto the column space of  $Z_{*,(m+1)}$  and  $\mathbf{e} = (\mathbf{0}_{m-K_{(m)}}^T, 1)^T$ . Define  $\mathcal{X}_{(m+1)}$  as in (A.32) and the same jointly asymptotic bound can be concluded for  $\mathcal{T}_{K_0,(m)}$  and  $\mathcal{T}_{K_0,(m+1)}$ , as in the previous case.

These construction steps can be implemented recursively, thereby completing the proof of Theorem 3.4.

**Acknowledgments.** We are grateful to the Associate Editor and two referees for their helpful comments and suggestions.

## REFERENCES

- AMATO, U., ANTONIADIS, A. and DE FEIS, I. (2006). Dimension reduction in functional regression with applications. *Comput. Statist. Data Anal.* **50** 2422–2446. [MR2225577](#)
- ASH, R. B. and GARDNER, M. F. (1975). *Topics in Stochastic Processes*. Academic Press, New York. [MR0448463](#)
- CAI, T. and HALL, P. (2006). Prediction in functional linear regression. *Ann. Statist.* **34** 2159–2179. [MR2291496](#)
- CAMBANIS, S., HUANG, S. and SIMONS, G. (1981). On the theory of elliptically contoured distributions. *J. Multivariate Anal.* **11** 368–385.
- CARDOT, H., FERRATY, F. and SARDA, P. (2003). Spline estimators for the functional linear model. *Statist. Sinica* **13** 571–591. [MR1997162](#)
- CARDOT, H. and SARDA, P. (2005). Estimation in generalized linear models for functional data via penalized likelihood. *J. Multivariate Anal.* **92** 24–41. [MR2102242](#)
- CARROLL, R. J. and LI, K. C. (1992). Errors in variables for nonlinear regression: Dimension reduction and data visualization. *J. Amer. Statist. Assoc.* **87** 1040–1050.
- COOK, D. R. and WEISBERG, S. (1991). Comments on “Sliced Inverse Regression for Dimension Reduction,” by K. C. Li. *J. Amer. Statist. Assoc.* **86** 328–332. [MR1137117](#)
- COOK, D. R. (1998). *Regression Graphics*. Wiley, New York. [MR1645673](#)
- CRAMBES, C., KNEIP, A. and SARDA, P. (2009). Smoothing spline estimators for functional linear regression. *Ann. Statist.* **37** 35–72. [MR2488344](#)
- DAUXOIS, J., POUSSE, A. and ROMAIN, Y. (1982). Asymptotic theory for the principal component analysis of a vector of random function: Some application to statistical inference. *J. Multivariate Anal.* **12** 136–154. [MR0650934](#)
- EATON, M. L. and TYLER, D. (1994). The asymptotic distribution of singular values with application to canonical correlations and correspondence analysis. *J. Multivariate Anal.* **50** 238–264. [MR1293045](#)
- EUBANK, R. and HSING, T. (2010). *The Essentials of Functional Data Analysis*. Unpublished manuscript. Dept. Statistics, Univ. Michigan.
- FAN, J. and LIN, S.-K. (1998). Test of significance when data are curves. *J. Amer. Statist. Assoc.* **93** 1007–1021. [MR1649196](#)
- FERRÉ, L. and YAO, A. (2003). Functional sliced inverse regression analysis. *Statistics* **37** 475–488. [MR2022235](#)
- FERRÉ, L. and YAO, A. (2005). Smoothed functional sliced inverse regression. *Statist. Sinica* **15** 665–685. [MR2233905](#)
- FERRÉ, L. and YAO, A. (2007). Reply to the paper “A note on smoothed functional inverse regression,” by L. Forzani and R. D. Cook. *Statist. Sinica* **17** 1683–1687. [MR2413540](#)
- FORZANI, L. and COOK, R. D. (2007). A note on smoothed functional inverse regression. *Statist. Sinica* **17** 1677–1681. [MR2413539](#)
- GU, C. (2002). *Smoothing Spline ANOVA Models*. Springer, New York. [MR1876599](#)
- HALL, P. and HOSSEINI-NASAB, M. (2006). On properties of functional principal components analysis. *J. R. Stat. Soc. Ser. B Stat. Methodol.* **68** 109–126. [MR2212577](#)
- HALL, P., MÜLLER, H. and WANG, J. (2006). Properties of principal component methods for functional and longitudinal data analysis. *Ann. Statist.* **34** 1493–1517. [MR2278365](#)
- HASTIE, T. J. and TIBSHIRANI, R. J. (1990). *Generalized Additive Models*. Chapman and Hall, New York. [MR1082147](#)
- HSING, T. and REN, H. (2009). An RKHS formulation of the inverse regression dimension reduction problem. *Ann. Statist.* **37** 726–755. [MR2502649](#)
- JAMES, G. A. and SILVERMAN, B. W. (2005). Functional adaptive model estimation. *J. Amer. Statist. Assoc.* **100** 565–576. [MR2160560](#)

- LI, K. C. (1991). Sliced inverse regression for dimension reduction. *J. Amer. Statist. Assoc.* **86** 316–327. [MR1137117](#)
- LI, Y. (2007). A note on Hilbertian elliptically contoured distribution. Unpublished manuscript, Dept. Statistics, Univ. Georgia.
- MÜLLER, H. G. and STADTMÜLLER, U. (2005). Generalized functional linear models. *Ann. Statist.* **33** 774–805. [MR2163159](#)
- RAMSAY, J. O. and SILVERMAN, B. W. (2005). *Functional Data Analysis*, 2nd ed. Springer, New York. [MR2168993](#)
- SCHOENBERG, I. J. (1938). Metric spaces and completely monotone functions. *Ann. Math.* **39** 811–841.
- SCHOTT, J. R. (1994). Determining the dimensionality in sliced inverse regression. *J. Amer. Statist. Assoc.* **89** 141–148. [MR1266291](#)
- SPRUIELL, M. C. (2007). Asymptotic distribution of coordinates on high dimensional spheres. *Electron. Comm. Probab.* **12** 234–247. [MR2335894](#)
- THODBERG, H. H. (1996). A review of Bayesian neural networks with an application to near infrared spectroscopy. *IEEE Transactions on Neural Network* **7** 56–72.
- XIA, Y., TONG, H., LI, W. K. and ZHU, L.-X. (2002). An adaptive estimation of dimension reduction space (with discussion). *J. R. Stat. Soc. Ser. B Stat. Methodol.* **64** 363–410. [MR1924297](#)
- ZHU, Y. and ZENG, P. (2006). Fourier methods for estimating the central subspace and the central mean subspace in regression. *J. Amer. Statist. Assoc.* **101** 1638–1651. [MR2279485](#)
- ZHU, Y. and ZENG, P. (2008). An integral transform method for estimating the central mean and central subspace. *J. Multivariate Anal.* **101** 271–290. [MR2557633](#)
- ZHANG, J.-T. and CHEN, J. (2007). Statistical inferences for functional data. *Ann. Statist.* **35** 1052–1079. [MR2341698](#)
- ZHONG, W., ZENG, P., MA, P., LIU, J. and ZHU, Y. (2005). RSIR: Regularized sliced inverse regression for motif discovery. *Bioinformatics* **21** 4169–4175.

DEPARTMENT OF STATISTICS  
UNIVERSITY OF GEORGIA  
ATHENS, GEORGIA 30602-7952  
USA  
E-MAIL: [yehuali@uga.edu](mailto:yehuali@uga.edu)

DEPARTMENT OF STATISTICS  
UNIVERSITY OF MICHIGAN  
ANN ARBOR, MICHIGAN 48109-1107  
USA  
E-MAIL: [thsing@umich.edu](mailto:thsing@umich.edu)